ECON INCORPORATED



# A PRELIMINARY STUDY OF THE STATISTICAL ANALYSES AND SAMPLING STRATEGIES ASSOCIATED WITH THE INTEGRATION OF REMOTE SENSING CAPABILITIES INTO THE CURRENT AGRICULTURAL CROP FORECASTING SYSTEM

INCORPORATED

Report #75-127-1
NINE HUNDRED STATE ROAD
PRINCETON, NEW JERSEY 08540
609 924-8778

# A PRELIMINARY STUDY OF THE STATISTICAL ANALYSES AND SAMPLING STRATEGIES ASSOCIATED WITH THE INTEGRATION OF REMOTE SENSING CAPABILITIES INTO THE CURRENT AGRICULTURAL CROP FORECASTING SYSTEM

Prepared for the
Office of Applications
National Aeronautics and Space Administration
Contract No. NASW-2558



June 30, 1975

## ABSTRACT

Remote sensing of agricultural croplands has been experimentally applied to the estimation of regional crop production statistics. Increasing accuracy and timeliness of the crop acreage component by remote sensing appears to be a major source of benefits from the new technology. Extending the crop survey application from small experimental regions to state and national levels requires that a sample of agricultural fields be chosen for remote sensing of crop acreage, and that a statistical estimate be formulated with measurable characteristics. The critical requirements for the success of the application are reviewed in this report. The problem of sampling in the presence of cloud cover is discussed. Integration of remotely sensed information about crops into current agricultural crop forecasting systems is treated on the basis of the USDA multiple frame survey concepts, with an assumed addition of a new frame derived from remote sensing. Evolution of a crop forecasting system which utilizes LANDSAT and future remote sensing systems is projected for the 1975-1990 time frame in this preliminary study.

## NOTE OF TRANSMITTAL

This report on a preliminary study of statistical integration of remotely sensed crop data into existing crop survey systems is prepared for the Office of Applications, National Aeronautics and Space Administration under Contract NASW-2558. It is based on a review of the current state of the art in remote sensing applications in agriculture and current crop survey methods. This study is entirely independent of the case studies in crop survey applications which are reported in other volumes under NASW-2558. The conclusions of the authors are their own, and do not necessarily reflect the detailed results of the economic models reported separately in the other volumes.

Submitted by: Francis Sand
Francis Sand

and: Robert J. Christie
Robert Christie

Approved by : George A. Hazelrigg, Jr.
George Hazelrigg
Project Manager

and: Bernard Miller
Bernard P. Miller
Project Director

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

# LIST OF FIGURES AND TABLES

# 1. ISSUES

## 1.1 Introduction

In repeated experiments, investigators have successfully applied LANDSAT multi-spectral digital data to the classification of crop acreage in various narrowly defined agricultural land areas. The idea that this application of LANDSAT holds promise for inventorying crop production on a large scale gained increasing support over the past several years. At present the Large Area Crop Inventory Experiment (LACIE) is testing this idea on a continental scale with the goal of a 90% accurate crop production estimate at the 90% confidence level for selected major crops. In order to pass from experimental verification to an operational crop survey system incorporating the use of LANDSAT multi-spectral scanner (MSS) digital data, it is essential to plan the linkage of these data with other crop data currently available from the US Department of Agriculture (USDA) crop surveys, and with meteorological data for processing yield estimates. The purpose of this report is to examine the requirements for integrating LANDSAT data into USDA crop surveys to further the aim of achieving an improved crop survey system.

Reviewing the investigations completed to date, we find that only the acreage measurement component of crop production estimates has been adequately developed in LANDSAT ex-

periments to date, to permit system design consideration for the integration of satellite and ground data to full-scale crop inventories. Accordingly, the main part of this discussion is limited to crop acreage measurements.

An open question is: Can LANDSAT data be used independently of USDA crop survey data to prepare national and state-level crop acreage estimates with acceptable accuracy? While this question is an issue of many ERTS investigations and experiments, the development of an improved USDA crop survey based upon satellite data integrated into a crop survey system would, in any case, be a necessary step in a well planned development program of the LANDSAT crop survey application. Thus, this report addresses the task of integrating LANDSAT data into the existing USDA crop surveys. Other tasks relating LANDSAT agricultural applications to more distant goals, including independent yield estimation from satellite data and/or global crop surveys, are also discussed briefly. However, it is not possible here to do more than indicate feasible scenarios for this later stage of developing a comprehensive worldwide satellite capability in agricultural surveys.

## 1.2 Sampling Strategy

While LANDSAT observations might ultimately cover most of the surface of the earth as the satellite sweeps through its 18-day cycle, the use of a complete census of agricultural

areas in crop inventories would be unnecessarily expensive. The objectives of a crop survey are:

- to provide timely and accurate data on crop planting, growth and harvesting
- to permit statistical estimates of crop production to be made within acceptable confidence limits.

Satisfying these objectives subsequently yields economic benefits through the publication of crop reports containing crop data and production estimates (or forecasts). It follows that a cost-effective approach to the LANDSAT crop survey application requires sampling the total crop area - or equivalently selecting sample segments from the agricultural land area observed by LANDSAT - for subsequent processing into crop acreage and yield information.

USDA crop surveys use two basic kinds of samples in the current Statistical Reporting Service (SRS) procedures for collecting crop data. One, a probability sample, is based on area segments selected from aerial photographs of the farmlands. Complete and objective crop data are obtained within the sampled segments by enumerators. The other kind, a non-probability sample, is obtained by mailing questionnaires to farmers on a carefully compiled list at certain fixed times of year. Those farmers on the list who do respond, supply much detailed and valuable crop and livestock information - which, however, cannot be checked, and thus is

more or less subjective. The total sampling is believed to represent 0.6% of the farmlands (by area) in the United States. Thus sampling error - the statistical variation between different samples - is a major contribution to the total error in the final estimates of crop production.

LANDSAT coverage of croplands is so extensive that the sampled area could, in principle, be extended to almost any desired fraction of the total area. In practice however, there are important considerations which limit this area fraction to some figure less than 100%, although substantially larger than 0.6%:

- The presence of cloud cover reduces the sample size obtained in any particular timespan by LANDSAT.
- The processing costs per LANDSAT frame are likely to be high, at least for early systems, so that the total acreage sampled must be kept to a modest level.
- The recommended approach toward development of the LANDSAT crop survey capability is evolutionary. Adjustments and refinements are easier to perform on smaller scale systems.

The design of the sample must be prepared by statisticians for efficient estimation of crop acreage within targeted confidence limits. In cases of mixed agricultural areas,

multicrop sample design is preferrable for reasons of efficiency. The sample should be stratified, with strata chosen to represent known intensities of agricultural activity and convenient political boundaries as with the USDA crop reporting districts (CRD's). Provision must be made for the rejection of sample segments after data acquisition, either because the cloud cover obscures essential data such as training sites, or because there are system-caused data losses in the segments. Then resampling these segments, or adjustment of the weights used for the surviving sample segments in the estimation formula will be required.

Statistical estimation of crop acreage from the sample requires "expansion" of the crop acreage measured in the sample segments containing the crop to the regional, state or national reporting level. Inferences made along scientific lines carry a known confidence, and thus are useful for resource managers seeking information about the crop. The final statistical estimate of crop acreage should supply reliable, accurate information to be integrated with other crop survey data at the appropriate level for the publication of crop production reports.

Sources of statistical variability in MSS data obtained by LANDSAT* include the time of year, the degree of cloudiness, the crop planting schedule, and the sample design.

---

*Assuming continuation of the present sun-synchronous orbit, sun angle is not a significant source of variability at a particular time of year.

In development of the crop classification and acreage mensuration techniques, there are many statistical inference problems to be solved. These problems are conceptually distinct from the subject of this section which concerns the design of an area sample for selected crops and the estimation of regional, state or U.S. crop acreage from the sample. Nevertheless, due to the complex nature of the data analysis, it may be found convenient to combine statistical inference problems at all levels from the pixel to the final large area estimate. This approach is not in any way precluded by the discussions of this section. There is, on the other hand, no necessity to attempt the linkage at this time.

## 1.3 Evolutionary Approaches

One approach to the development of a new technology application such as remote sensing of agricultural crops is to implement parallel systems (of crop forecasting) with the intention of phasing out the less efficient system as soon as possible. Another approach, which we recommend here, is to use the new technology in conjunction with the existing system, effecting a gradual integration of new and old data collection and analytic techniques.

There is, at present, insufficient experience in applying LANDSAT data to crop surveys for an integrated satellite-aircraft-ground truth crop inventory system to be fully and accurately specified. Yet the positive evidence accumulated to

date allows for a reasonable expectation that the LACIE and principal investigator results will lead to a first-generation operational crop acreage estimation system, at least for some crops and some geographic regions. The successful integration of the LANDSAT crop information into existing USDA/SRS procedures, requires that the system development should proceed in an evolutionary manner in spite of some apparently revolutionary aspects of the LANDSAT capability in agriculture. The use of LANDSAT data to estimate leaf area index (LAI) or other yield correlatives may become significant one day, and thus be acceptable as a useful addition to the existing USDA yield measurement programs. But so long as the degree of correlation is still very weak, it is necessary to continue using the existing methods without modification, while at the same time implementing LANDSAT-based changes in the acreage measurement programs.

In order to achieve a fundamental change in agricultural crop reporting accuracy and comprehensiveness, it will undoubtedly be important to achieve a meaningful articulation between LANDSAT measurements of crop acreage and USDA/SRS data handling. This imposes, at the very least, a requirement for a LANDSAT acreage reporting format which can be directly utilized by SRS together with its other multiple-frame area surveys. Timing of reports will also have to be considered.

The evolutionary approach to the subject provides for incremental steps to be taken which supply new crop survey information to SRS only after thorough testing and demonstration of the reliability of that information, and after agreement has been reached with SRS regarding the format of the information.

## 1.4 Economic Issues

It has been determined by detailed economic analysis that substantial benefits could be obtained by U.S. food consumers as a result of improvements in crop production forecast accuracy. The specific magnitude of the benefits has been obtained in a concurrent ECON study* as a function of three parameters of the total information system: the planted acreage estimation accuracy, the frequency of measurement of planted acreage and the data lag between the time of the measurement and the issuance of a production (forecast) report. Clearly, the values that these parameters take on are a function of both the satellite system and the ground processing system. (Measurement of the yield component of production is assumed to continue at the current degree of accuracy.) The magnitude of the disbenefits associated with errors in current USDA crop production forecasts are estimated to be $211 million for wheat and $40 million for soybeans annually. Hence, even small

*The Value of Domestic Production Information in Consumption Rate Determination for Wheat, Soybeans, and Small Grains, ECON, Inc., Report No. 75-127-3, Princeton, N.J., August 31, 1975.

improvements in these forecasts could provide sufficient benefits to justify development of the new space-based capabilities required.

The proposed application of LANDSAT agricultural crop surveys requires implementation on a national (for U.S. Crops) and, perhaps, worldwide scale. The benefits estimated do not accrue if the results of the application are not integrated into a crop production reporting and disseminated on a non-discriminatory basis. Inasmuch as the application of LANDSAT data provides reliable acreage estimates only (at least initially - good yield estimates may follow later) there is no economic basis at present for considering distribution of the LANDSAT data other than through a statistical reporting service which has the necessary capability to integrate LANDSAT data with the other elements of crop production estimates in order to obtain improved production estimates and forecasts.

In addition to the benefits associated with improved crop production estimates and forecasts obtained on a national scale, additional benefits would result from the dissemination of local and regional statistics, for example, at the state or county levels. Provisions for this secondary distribution can also be made through the statistical reporting service responsible for the national statistics as much of the necessary machinery already exists for cooperation between the various concerned agricultural agencies. The development of marketable

information products from the crop survey application can reasonably be anticipated at present and is, of course, an important economic issue; however, until the development of practical LANDSAT data processing techniques is further along and such products forthcoming, we envision that the new information would be used mainly by governmental agencies responsible for assessing crop production quantities and crop conditions. To wait for the growth of private enterprises to process LANDSAT data into marketable products might entail considerable loss of time, during which benefits from a LANDSAT capability in agriculture could have been realized through the public sector. On the other hand, the growth of a market for specialized information products and services derived from satellite images of agricultural crop and rangeland may be expected to occur concurrently with the improvement of the national and state crop production estimates and this market will undoubtedly be served partly or wholly by private enterprises under the existing system for the distribution and pricing of LANDSAT data.

The above issues notwithstanding, the major economic issue concerning the implementation of LANDSAT data into crop production estimates and forecasts deals with the present uncertainty in the technical capability that a LANDSAT type satellite-based system might offer and what computational (automatic) and manual treatment of the data are necessary to achieve this capability. To be sure, the implementation of LANDSAT data

into a crop production estimation and forecasting system is still very much a topic of research, despite the fact that some rather definitive statements might presently be made regarding the interim, if not the ultimate, system capability. Thus, the problem of an implementation schedule becomes quite important. Should a system of lower capability be implemented early as opposed to a system of higher capability delayed in time? To what extent should the system rely on manual versus automatic processing vis a vis area coverage, data lag and flexibility for system growth? Should the initial capability be optimized, for example, to produce the maximum net economic benefit or should the system be designed merely to meet certain institutional goals while allowing for added freedom of growth? A substantial policy analysis should be addressed to the potential implementation scenarios. This analysis should include a detailed cost and capability analysis of the alternatives and an analysis of the risks associated with each alternative. The remainder of this report sets the stage for such a study.

# 2. THE INTERFACE BETWEEN LANDSAT AND USDA/SRS

## 2.1 The Interface in 1975 - 1980

From the multiplicity of uses of LANDSAT imagery reported in the scientific literature, there would appear to be a bewildering array of choices for the organization of the crop survey applications. However, there are pertinent facts concerning the economics of the applications which narrow the field of choice. In order to develop the applications in economically viable ways, there are several prerequisites that must be satisfied. We consider the primary requirements to be:

- The survey should measure economically important aspects of agriculture, such as crop production for a major crop at state or national levels.
- The results of the survey should be available to all interested users in the agricultural community in a timely fashion.
- The format of the information developed from LANDSAT data should be acceptable to the users, which implies that the presentation should be relatively effortless to interpret.
- The processing of LANDSAT data should be done efficiently to avoid excessive costs or delays.
- The statistical nature of crop survey information requires a scientific application of statistical techniques to ensure accuracy and high confidence in the information.

Within the guidelines of these constraining considerations, the major choices for the crop survey applications appear to be encompassed within the following questions:

1. Which crops to survey?
2. Which geographic areas to cover and how complete does each coverage need to be?
3. What are the crop measurements (statistics ) to derive from LANDSAT data?
4. Who are the end users of the processed results?
5. To what extent should the applications be locked in to existing institutional procedures for publishing agricultural information?

For the purposes of this preliminary study of the integration of LANDSAT applications with USDA/SRS procedures, the scope of our inquiry is further narrowed to an examination of the data handling and statistical problems under the following assumptions:

- The responsible user agency will receive computer compatible tapes of geometrically and radiometrically corrected LANDSAT data, or that agency will have in-house capability to perform these preprocessing corrections.
- The agricultural community will receive improved crop production estimates and forecasts as a result of the use of LANDSAT data in the crop surveys.
- The use of LANDSAT data for the crop surveys will be

efficiently organized, so as to avoid unnecessary errors of interpretation, delays and costs in processing the data, and to facilitate the achievement of the desired goals in the improvement of crop production forecasting.

The interface can now be characterized through an analysis of USDA/SRS crop production estimation procedures, together with a review of results of principal investigations using LANDSAT data to classify and measure crop statistics. The crop production estimate is a product of two components: acreage harvested and yield per acre. These are sampled, measured and estimated in separate programs by SRS.

Most likely, early systems will be built to obtain improvements in crop acreage estimates until such time as crop yield estimation can be significantly enhanced through remote sensing of the crop. We will review the interface issue under acreage and yield headings separately.

### 2.1.1 Acreage Interface

The Agriculture Handbook No. 365* published by USDA refers to the acreage estimates in the following terms:

> "In general, the progression of acreage estimates is from prospective plantings to acreage intended for harvest to acreage actually harvested. Most spring-sown field crops follow this sequence: (1)

*"Major Statistical Series of the U.S. Department of Agriculture," Vol. 8, May 1971.

acreage intended for planting as of March 1, released about mid-March; (2) acreage planted and acreage for harvest, released with the midsummer report; and (3) acreage planted and harvested, released in the December Annual Crop Production Summary. Fall-sown rye and winter wheat depart from this sequence, with seeded acreage estimated in December of the year preceding harvest, and winter wheat acreage for harvest in May of the next year."

"The total harvested acreage of many crops is broken down into utilization groups. For example, although the major use of corn and sorghum is for grain, separate estimates are also made for the acreage harvested for silage and for forage, including acreage grazed or hogged."

"In general, acreage estimates are based on two types of information: (1) acreage data for a given crop season, obtained from the quinquennial census of agriculture, state farm censuses, or some other complete or nearly complete enumeration; and (2) indicated acreages obtained by questionaires from samples of farms or processing plants."

"Major national surveys to collect data on acreages of field crops and some seeds and vegetables are conducted annually around March 1, June 1, and during the fall. The March survey is in large measure a nonprobability mail survey, whereas the June and fall surveys are based upon both mail and probability samples. Acreage utilization and production data are also obtained for a number of major crops on the fall survey."

From our point of view, an important feature of the ESDA/SRS methodooogy is the use of multiple-frame sampling. Part of the sample used to prepare crop acreage estimates for major crops is obtained from the list frame, the other part from an area frame. The latter is a probability sample, while the former is not. There are some farms which are unavoidably included in both frames. Provided that the overlap portion is

identified this does not cause any problems of estimation. The expansion of the overlap portion of the sample must be undertaken separately to give the proper weights to these units. In addition to being multiple-frame, the survey design is at the same time stratified. The stratification is obtained by dividing each state into strata according to intensity of agriculture; then each stratum is futher subdivided into sampling units of variable size (about one square mile for very intensely cultivated land).

USDA uses aerial photography to construct area frames. The photographs are updated on approximately a 5-year cycle. Recently, the use of LANDSAT data has been proposed in the framing of the area sample.* Clear delineation of boundaries of fields is necessary in constructing the area frame so that enumerators can identify these fields on the ground correctly. The USDA evaluation of this application of remote sensing is expressed in "Scope and Methods of the Statistical Reporting Service," Miscellaneous Publication No. 1308.

It is evident from the work of principal investigators in the agricultural crop survey applications area that LANDSAT data can be used to construct independent acreage estimates for some crops, such as winter wheat, given the necessary amount of "training" data for the correct identification

---

*Crop Identification and Acreage Measurement Utilizing ERTS Imagery, William H. Wigton and Donald H. Von Steen in: Third ERTS-1 Symposium (Dec. 1973) pp. 87-92.

of the crop by the classifier system. Further research on the classification of agricultural crop areas from LANDSAT data is progressing, and it is not unreasonable to expect that the capability to classify most of the major crops correctly from cloud-free LANDSAT frames will be proven in the near future. This capability might require repeated "looks" at the crop-growing area to achieve an acceptable level of crop classification accuracy. The use of spectral signatures to classify crops and pixel counts to mensurate crop acreage is clearly a different technology when compared with the current USDA program for acreage estimation. In what way can this new technology be used most cost-effectively to supplement and improve the USDA acreage estimates? The interface, as it can be defined today, is bounded on the one side by the statutory requirements for the Crop Reporting Board to report timely and accurate crop production figures at specified times within a limited budget; on the other hand by the uncertainties and unresolved issues concerning the application of remote sensing techniques using LANDSAT data to large area crop inventorying.

In the research environment, where timeliness is not a major factor, high accuracies have been reported for LANDSAT-based independent crop acreage estimates within narrowly defined limits of cartographic area and time of year.* These findings

*See, for example, Agricultural Inventory Capabilities of Machine Processed LANDSAT Digital Data by Dietrick, Egbert and Fries at NASA Earth Resources Survey Symposium, June 1975 (Houston).

relate to few crops and are not yet extended to statewide or national crop acreage estimates. Whether it is feasible to do so with the existing technology is still an open question. The promising aspects of the LANDSAT application appear to reside in the following points:

- LANDSAT possesses the capability to supply multispectral images of a very large agricultural area in a short span of time.
- LANDSAT data are objective.
- LANDSAT data are usually current, within the crop cycle of the year of study, subject to cloud-free scenes being obtained.
- LANDSAT images will be most likely amenable to automatic interpretation and, through advanced processing techniques, will most likely generate cropacreage estimates of high accuracy within a short timespan after data acquisition.

All of these considerations provide justification for an intensive effort to do research and develop cost-effective techniques for using LANDSAT data in the crop acreage estimation program of the United States, either through direct utilization by USDA or by another Federal agency acting in concert with USDA. The interface itself can be more sharply defined only by pursuing such investigations. A valuable beginning is found in the LACIE effort, which will undoubtedly

reveal further promising achievements and, perhaps, also limitations to the scope of LANDSAT applications to crop surveys.

### 2.1.2 Yield Integration

The USDA yield program is described briefly as follows by the Agriculture Handbook No. 365: *

YIELD AND PRODUCTION

"Yield refers to production per acre measured in units such as pounds, bushels, hundredweight, and so on, whereas production relates to total units produced. Forecasts and estimates of yields and quantities produced for crops are usually provided as of the first of each month during the growing season. The preponderance of the forecasts and estimates fall within the period July 1 to December 1, but for crops not in season during this period, primarily vegetables, estimates are timed appropriately."

"Forecasts and estimates are two distinct concepts. Forecasts refer explicitly to expectations of what is likely to be accomplished at some time in the future, such as a prediction of the yield or production of an immature crop. Estimates generally refer to a measure of accomplished fact, such as crop production at or after harvesttime."

"It should be clearly understood that a forecast is a statement or report of the prospective yield or production, on the basis of known facts on a given date, assuming weather conditions and damage from insects or other pests during the remainder of the growing season will be about the same as the average of previous years. Potential based on current conditions may be appraised accurately, but if weather or other conditions change, the actual outturn may differ somewhat from the forecast. As a crop develops, crop reporters periodically submit appraisals of probable yield or production on their farms and in their localities, and the averages of these reported data are translated into forecasts by the Crop Reporting Board."

"Monthly forecasts and end-of-year estimates for several crops in many States are also based on objec-

---

*"Major Statistical Series of the U.S. Department of Agriculture," Vol. 8, May 1971.

tive yield survey data. In the objective yield surveys, trained enumerators visit selected fields and orchards chosen on a probability basis to make counts and measurements of plants and fruit characteristics on small plots located in sample fields or in sample trees. This is done during the growing season for indications of the probable final yield when the crop is mature and harvested. At harvest time actual yields in the sample plots are measured, and sample plots are gleaned after harvest to measure harvesting losses. From these sample results, forecasts and actual yields are computed along with sampling errors and these are made available to the Crop Reporting Board for making estimates."

"When final survey indications and all check data for a crop become available, usually some months after completion of harvest, the official estimates of production are reviewed and revised, if necessary. Annual revisions are scheduled in advance and are released at essentially the same time every year."

The determination of the expected yield per acre, even for such a widely studied crop as wheat, is a complex and difficult task. There are numerous factors affecting plant growth, and the use of models to obtain regional (state) or specific (local) predictions of yield is far from being perfected. For a detailed review of the issues we refer to the Goddard Task Force on Agricultural Forecasting (GTFAF),* selections from which are reproduced in the Appendix to this report. Some of the difficulties relate to the complexity of the relationship between yield and the crop growth factors. Other difficulties are met in the data collection area. Meteorological data, already being collected by satellites, can provide

*The Use of the Earth Resources Technology Satellite (ERTS) for Crop Production Forecasts, Draft Final Report, Task Force on Agricultural Forecasting, edited by D.B. Wood, NASA Goddard Space Flight Center, July 24, 1974.

some inputs for AGROMET yield determination models (see GTFAF). It appears likely that crop stress factors which limit yield can be detected* and measured by LANDSAT. Further assistance to the yield program may be provided from LANDSAT images by detection of crop abandonment. According to our literature survey, to date no demonstration has been made of a capability to measure the yield per acre of a crop from high-altitude remote sensing data. Numerous studies indicate that valuable inputs to yield estimation models may be obtainable from satellites, particularly the weather satellites, but also including LANDSAT. For the present purpose, the interface must be characterized by those factors, related to yield, which are partially or wholly measurable by analysis of LANDSAT data.

## 2.2 THE INTERFACE IN 1980-1990

Anticipating the evolution of a satellite-based remote-sensing applications system for crop surveys, in the manner described previously (Section 1.3), there is a different perspective of the interface. If one postulates an operational system for automatic classification of agricultural crops in all geographic units of the United States from satellite remotely sensed data, with the concomitant acreage mensuration of high accuracy, available on a 24-48-hour basis, the user

*Wheat: Its Growth and Disease Severity as Deduced From ERTS-1, E.T. Kanemasu, C.L. Niblett, H. Manges, D. Lenhart, M.A. Newman in Remote Sensing of Environment 3, 255-260 (1974).

agency would be able to use this information to replace older and less cost-effective survey techniques, as well as to derive new information products at the local level. While we hesitate to predict which techniques might be replaced or which new products created, the conclusion, as far as the interface is concerned, must be that such a system could become an integral part of the crop surveys after 1980, rather than a superficial addition to the multiframe survey system of today.

Beyond integrating acreage estimation data from LANDSAT and successor systems into the crop surveys, there remains a host of potential applications which may provide early warning information on crop conditions or survey information on other aspects of agricultural activity. These applications would need to be handled individually with due consideration for user demand and institutional charter, but we will not attempt to pursue the topic any further than that. Some of them may prove suitable for commercial exploitation; others may require new agency arrangements; still others may fit into the organizational framework of existing agencies such as USDA/SRS.

# 3. CRITICAL REQUIREMENTS OF THE INTEGRATION

## 3.1 Sample Design

The sample of area segments within the U.S. agricultural lands which are to be registered, classified and measured by processing LANDSAT data can be considered as a mechanism for selecting a manageable portion of the vast amount of data acquired. Processing of all relevant* data in a timely and cost-effective way is an option to be evaluated. This provides a census of the agricultural land, but it is not a total census in that some areas will be excluded by cloud cover, and fields which are too small for the classifier are also lost. A scientifically designed statistical sample of the agricultural land is an alternative option which is likely to prove cost-effective. Design criteria of the sample, which should be taken into account are:

(1) the size of the region for which the sample is intended: U.S. nation, 48 coterminous states, one state, crop reporting district, county, etc.,

(2) the intensity of agricultural activity relating to the crops of interest,

(3) the probability of obtaining a cloud-free LANDSAT frame, or sufficient cloud-free area within

---

*Obviously data pertaining to cities, mountains, lakes, deserts, etc. can be excluded.

the frame,

(4) the number of LANDSAT passes which must be used to construct the sample,

(5) the acceptable level of sampling error,

(6) the need for training sites for the classifier within the sample segments.

Some of these points, such as (1) and (5), relate to objectives of the survey. Others, such as (2) and (3), relate to the physical state of the region and its atmosphere at the time of the survey. The remainder, (4) and (6), relate to the techniques used for registration, classification and mensuration of crop acreages. Each of the issues - survey objectives, physical state of the environment and measurement techniques - must be resolved fully at the time of survey implementation.

One of the technical issues to be resolved concerns the use of agricultural fields as an integral part of the acreage classification and measurement processing of LANDSAT images. The choice of technique in this area has some bearing on sample design since efficient sampling and estimation would require knowledge of field size distribution if fields are used as a structural basis for crop classification. The following table presents a brief overview of the comparative advantages of two methods.

The sample design itself can be undertaken without undue difficulty once the major issues outlined above have been

Table 3.1 Comparison of Remote Sensing Techniques for Crop Classification

| Field Classifier | Pixel Classifier |
|---|---|
| Lack of knowledge of field size distributions | Field size distributions not needed |
| Fields are useful for classification of crops in that they provide spatial context | Clustering of contiguous pixels can be done to a limited extent - some of the spatial context is lost |
| Reduction of database size by using fields | Simpler structure of a "coordinate grid" database |
| Variations within fields can lead to increased mensuration error if they are not fully accounted. (e.g., small ponds, bare patches, etc.) | Classification of isolated pixels may cause bias in estimates - fractional pixel classification is difficult |

resolved. Following accepted survey techniques, one would stratify the population with strata defined on the basis of known agricultural practices and crop calendars. Each stratum would contain, for instance, a geographically contiguous area containing a more or less known amount of activity relating to the crops of interest. The segments or sample units would be selected from within each stratum by one of two standard methods, sample size proportional to strata size (sampling the same fraction of each stratum) or optimal allocation, taking into account the variances of the measurements within strata and the "cost" of sampling if any differences occur between strata.

An additional criterion which might be employed in the sample design is cloud cover. The samples should be selected to increase the probability of obtaining cloud-free samples from areas which are frequently obscured by clouds, and these samples should be weighted to reflect the relative scarcity of cloud-free conditions. In order to do this it will clearly be necessary to develop database on regional cloud statistics for time of year. This issue will be reviewed in the next section of this report.

The total size of the sample will be determined by the economics of data acquisition and processing in relation to the objectives of the survey. If there is an institutional requirement to achieve a predetermined total error level, for example if the objectives of the LANDSAT application include obtaining a total crop acreage estimation error no larger than the currently existing value, then one may control the sampling error, $E_S$, in relation to the measurement error, $E_M$, to achieve this total error level:

$$E_T = \sqrt{(E_S^2 + E_M^2)}$$

## 3.2 Missing Data Due to Cloud Cover

Cloud cover can present a satellite remote-sensing applications system with a critical problem. In the case of a crop survey using sampling with fixed-area segments on a specified date, the presence of cloud cover causes loss of signif-

icant quantities of data, possibly all data on crops within the segments. If the sampling is timed to capture LANDSAT images of agricultural areas at a particular point of the crop cycle, this loss may severely reduce the overall quality of the sample. The results of crop acreage estimation derived from the sample may suffer from two forms of distortion due to cloud cover:

- The sample may be biased, due to the unrepresentative nature of the cloud-free portions of the sample for which data was actually obtained.
- The sample may result in too high a level of sampling error due to the effectual reduction in sample size by the cloud cover problem.

If it is possible within the time frame of the sampling procedures, repeat observations on a later date should be obtained to minimize these distortions. Otherwise, there are two main alternative "safeguards" against distortion due to cloud cover:

- A sample using "floating" rather than fixed area segments selected from the cloud-free portions of the images.
- A sample that is overdesigned so that a cloud-free subsample can be selected as necessary.

Neither of the alternative safeguards guarantees a satisfactory solution 100% of the time, although experience may show that one or both of them work well enough to provide statistically acceptable results. It is also clear that, from

LANDSAT survey data alone, crop acreage estimation for the smaller geographic units, e.g., counties, can be rendered infeasible by cloud cover if the data are narrowly limited in time. Whenever the data are obtained from several passes of LANDSAT the cloud cover problem is greatly reduced, and it is possible to calculate the minimum required number of passes to obtain a desired confidence level for the crop acreage estimate in each geographic unit. The nature of this critical problem is therefore one which allows solution only after the techniques of crop classification from the LANDSAT data have been formally specified. These technique specifications must be either:

- time-insensitive within a wide range of the crop growth cycle, or
- based on a sample design which explicitly recognizes the existence and geographical distribution of cloud cover at the time the sample is obtained.

For the latter purpose, a detailed study of cloud statistics would be required on a current basis for the time of year and geographic region of interest. The 1969 Study, "Cloud Statistics in Earth Resources Technology Satellite (ERTS) Mission Planning" by Vincent V. Salomonson provides seasonal frequencies of 30% or less cloudiness for the contiguous 48 states. Further detail would be required to design cropland samples which recognize cloud cover probabilities explicitly.

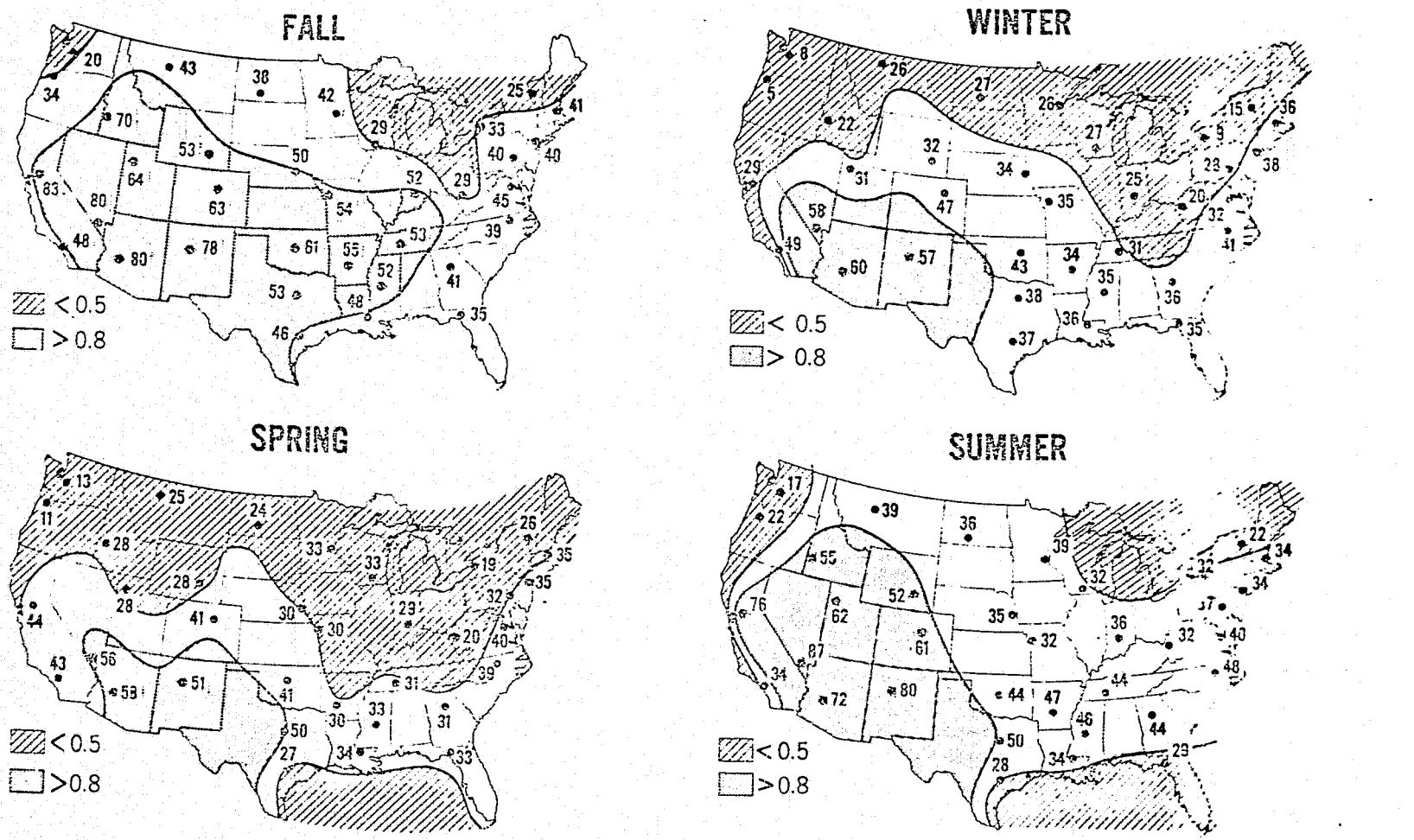



Figure 3.1 Four maps of the United States showing the frequency in percent of 30% or less cloudiness at 35 stations and the general locations where the probability is > 0.8, 0.5-0.8, and < 0.5 of seeing 30% or less cloudiness on at least 2 out of 5 passes during a season. The frequencies shown were compiled for the four seasons by Smith and Shafman (1968) and are based on ten years of record at each station.

Source: "Cloud Statistics in ERTS Mission Planning" by V. Salomonson, GSFC, 1969

## 3.3 Comparability of Satellite and Ground Survey Data

There are two major differences between remote sensing surveys of crops and conventional surveys employed by USDA.

(i) The sampling of farms or fields is based on totally different "frames."*

(ii) The timing of LANDSAT data acquisition is significantly different compared with the USDA conventional surveys.

We will deal with each of these separately in this section as applied to the estimation of crop acreage. Discussion of integrated yield programs presents far more difficult problems because of the complexity of the yield prediction models.

### 3.3.1 Different Sampling Frames

In one sense the difference in frames and sampling units between a LANDSAT survey of agricultural areas and a conventional enumeration or mail-out survey is no problem because the USDA already uses a multiple-frame approach. However, when one considers in detail the integration of the LANDSAT and conventional surveys, one is faced with a critical requirement:

- to statistically combine acreage from the LANDSAT

---

*We are not referring to LANDSAT image frames of 100 n.mi x 100 n.mi, but to the sampling frame which provides an operationally useful definition of the population to the statistician who must define the procedure by which samples are to be selected from the population.

data with acreage from the USDA enumerative and mail-out surveys, one must be able to specify how the LANDSAT acreage was sampled.

This requirement is not critical if:

(1) area segments are cartographically defined as a sampling frame, and

(2) LANDSAT images are registered with respect to those segments, and

(3) a probability sample of the segments is selected for crop acreage classification and mensuration. In this case, the results of the LANDSAT acreage survey can be statistically integrated with the results of the USDA enumerative surveys and mail-out surveys using standard techniques - essentially a weighted averaging procedure with the weights determined in relation to the standard errors of the estimates that are obtained from the several sources of information. However, if any one or more of the steps (1) - (3) outlined above are not followed, for any reason, then integrating the survey results may be difficult.

The estimates of crop acreage which might be obtained independently from LANDSAT data have different statistical characteristics from estimates derived by ground surveys. Apart

from the classification errors - such as confusion of similar crops - and the cloud cover problem, they differ substantially with regard to sampling errors. The total error of estimation derives from several sources, only one of which is sampling error. In USDA crop surveys based on enumeration of crop acreages within area segments, the measurement error is very low ( 0.5%), while the sampling error is much larger due to the small fraction of total area sampled. When LANDSAT data are processed for estimation of crop acreages, the measurement error becomes a combination of several factors and is likely to be larger than USDA enumerative crop surveys. On the other hand, the sampling error will be reduced because the fraction of croplands sampled can be substantially larger than existing surveys. Integration of LANDSAT data with USDA crop survey data should be planned to take advantage of one of the main virtues of LANDSAT images: their large area coverage. Needless to say, the information in independent estimates of crop acreage could be used in other ways to:

- check other survey results,
- develop new schedules of crop reporting,
- monitor progress in planting or harvesting.

From the economic studies of remote sensing satellites it does not appear that these other uses would be cost-effective by themselves. Once the system is developed for the agricultural crop survey mission, however, a list of minor applications become incrementally justifiable.

### 3.3.2 Different Timing of Acquisitions of Survey Data

The 18-day repeat cycle of each LANDSAT satellite permits, in principle, frequent updates of crop acreage estimates when compared with the reporting of crop data currently obtained by USDA. However, there are several factors which in practice will reduce the update frequency considerably:

- classification of crops from LANDSAT images with acceptable error levels may require multi-temporal data,
- several repeat observations of the same area may be needed to obtain sufficiently cloud-free scenes,
- some crops will only be identifiable or distinguishable from other crops at a particular time of year in LANDSAT images.

Perhaps the most positive statement that can be made about the LANDSAT frequency of data acquisition at the present time is that it provides an opportunity to obtain some crop acreage estimates on a monthly basis at state and perhaps even county levels. While these would not be complete, they would provide a new agricultural information service based on LANDSAT images. Whether or not these monthly regional crop acreage estimates would be immediately integrated with USDA/SRS preliminary survey results, or held until the completion of the annual crop survey, they would serve as a basis for improved

crop forecasting. The method of improvement would be either through independent preparation of new forecasts based on LANDSAT results, or through integration of those results with USDA crop survey data.

## 3.4 Uses of Ancillary Data in LANDSAT Applications to Crop Surveys

Due to the special nature of the LANDSAT image analysis procedures for classifying crops and mensurating crop acreage, there is a need to use considerable ancillary data to assist the classifier and to achieve maximum precision in the results. There is (potentially) a critical requirement in this matter due to the large amount of current agricultural data which the multi-spectral image analysis system would require. If one employs automatic (computerized) classification, which is considered essential for a cost-effective operational system, the ancillary data must be organized in a computer databank and retrievable by the classification programs. This will require a substantial amount of coding and input of the ancillary crop data to keep the data bank current and in general to maintain it in usable form. In summary: the planning and organization of a databank containing up-to-date agricultural crop information* with data such as local planting times will be a critical requirement for the integration of the LANDSAT crop survey applications with USDA crop surveys.

*See Appendix C for a discussion of the issues concerning the use of crop calendars to assist in the task of remote sensing identification of crops.

# 4. CONCLUSIONS & RECOMMENDATIONS

## 4.1 Conclusions

The use of LANDSAT in U.S. crop surveys has significant potential benefits if the accuracy and timeliness of existing crop production estimates can be improved significantly thereby. To achieve the goal, it is necessary that a qualified organization should receive the LANDSAT crop survey information and integrate it with crop information obtained by other methods. So long as LANDSAT supplies only the acreage component of a U.S. crop production estimate*, there is a substantial body of agricultural data which would be required in addition to LANDSAT data. At the present time, only the USDA has the independent capability to acquire, process and integrate all of these data into a timely and accurate crop report. The development of the remote sensing capability in agriculture into a crop reporting system requires expertise far beyond the classification and interpretation of LANDSAT data on crop producing areas. We feel that technological improvements in crop survey should be pursued in full cooperation with USDA and should have full support from existing USDA bureaus and institutions for preparation of crop reports in order to achieve maximum public acceptance and economic usefulness.

---

*Production=Acreage x Yield per acre

Progress in the development of automatic processing of LANDSAT data may lead eventually (say in the 1980's) to an independent, stand-alone system for crop reporting. However, even this conclusion is doubtful and based only on certain broad assumptions about the new technology rather than demonstrating facts.

In global crop surveys, the situation is more complicated due to

(1) the incompleteness and inaccuracy of much of the existing crop data for foreign countries, and

(2) the scale of the global survey task; complete and accurate crop reports for worldwide agriculture would require many times as much data processing as U.S. crop surveys.

Integration of LANDSAT data into foreign agricultural surveys should be pursued with the cooperation of USDA/FAS, while research is in progress to develop successful techniques to extract crop acreages and yield indicators from LANDSAT data. Obviously, much has to be learned before one can confidently predict a global crop survey capability using LANDSAT (or any of its successors) as the prime data source. We have concluded that the integration of satellite and ground data on worldwide crop production should be undertaken only after the successful demonstration of advanced interpretation techniques for remotely sensed data on agricultural areas outside the U.S. and Canada.

## 4.2 Specific Recommendations on Integration of Data

### 4.2.1 Techniques

Development of techniques to select, classify and mensurate a statistical sample of LANDSAT data on crop producing areas in the U.S. must be continued. Expansion of the sample results to provide an estimate of the crop production for the reporting region - whether that is county, state or nation - must be scientifically researched. In addition to the geographical considerations of sample design, the problems of timing of data and selection are critical, particularly in the presence of cloud cover.

We recommend that NASA should promote research on the following technical issues relating to the U.S. crop survey application of LANDSAT:

- overcoming cloud-cover problems on the sampling of relevant U.S. crop data from the LANDSAT data resource,
- the development and updating of the databank of "ancillary" agricultural data (i.e., not remotely sensed) is required for automatic processing of remotely sensed crop data into meaningful crop production estimates,
- the sampling of LANDSAT data for efficient statistical inference on national and regional

(state and county) crop production - stratified, multi-frame samples in relation to variety of cropping practice, and time of year are expected and

- the accurate cartographic registration of LANDSAT images to allow for easy comparability of the LANDSAT intepretive results with existing USDA crop survey results.

4.2.2 Evolutionary Approach to Integration

We take the position that there are advantages, both technical and economic, to an evolutionary staged approach to the integration of LANDSAT crop data into the crop reporting system. A possible scenario for this solution is presented for illustration of the method:

Stage IA

Develop statistical and data processing techniques for using LANDSAT data to obtain state and national crop acreage figures for a few selected crops in the United States.

Stage IB

Develop crop yield models and associated inputs for yield measurement from LANDSAT data. Explore the feasibility of obtaining an accurate crop yield measurement system using LANDSAT data to provide local crop condition data in each crop reporting district (CRD) or other regional subdivision.

Stage IIA

Use new crop acreage estimates from LANDSAT together with USDA crop survey data in an integrated crop reporting system.

Stage IIB

Develop an independent crop acreage reporting system for all major crops amenable to remote-sensing classification.

Stage IIIA

Develop new agricultural informational services based on daily, weekly or monthly regional surveys of crops from LANDSAT data e.g., planting progress reports (acreage), harvest progress report (acreage), crop stress warnings (yield factor), crop condition assessments during growing season (yield factor).

Stage IIIB

Develop a new crop survey system integrating fully the satellite data with ground data and replacing older, less cost-effective survey techniques with satellite remote sensing techniques.

The logical relationship between the stages is indicated in Figure 4.1. The branch ending at IIIA describes an integrated approach to the use of LANDSAT imagery for crop acreage estimation based on low accuracy of the LANDSAT crop survey results. The other branch refers to an independent LANDSAT-type system for crop survey based on high accuracy of survey results. Stage IB develops inputs to yield prediction models and is independent of the acreage developments.

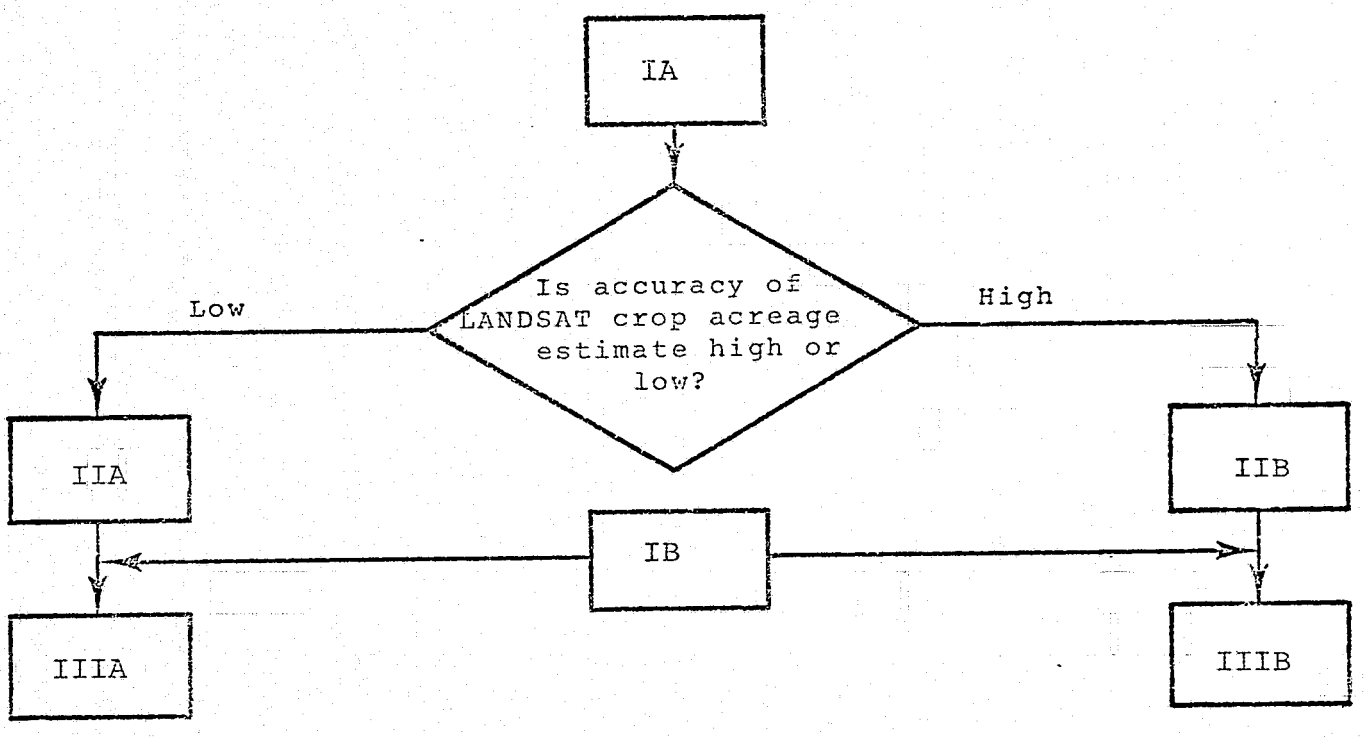



Figure 4.1 Logical Relationships between Evolutionary Stages

APPENDIX A

# THE MEASUREMENT OF THE YIELD COMPONENT*

## The Determinants of Wheat Yield

This section will provide an overview of the primary factors which impact upon yield. Later in this study we will illustrate which of these factors are contained in yield models.

The factors affecting plant growth are numerous and complex and their affects vary with the growth stages and the time of planting. Plant physiologists have defined more than a dozen stages in plant growth when observations and measurements can be made. Most of the literature consulted in this study referred to from six to nine stages. Two commonly used keys for wheat growth stages are

| | Growth Stage | Growth Stage |
|---|---|---|
| a. | Tillering | Seedling (emergence) |
| b. | Early Joint | Tillering (5 or < 5 leaves) |
| c. | Late Joint | Tillering (> 5 leaves) |
| d. | Boot | Jointing |
| e. | Heading | Boot |
| f. | Anthesis | Heading (50% of head out) |
| g. | Berry | Flowering |
| h. | Milk-Soft Dough | Dough |
| i. | Ripe | Ripe |

The stages which have been most widely used as growth parameters are emergence, heading and ripe. There is considerable year-to-year variation in the time of occurrence of each growth stage as well as the degree of plant development in each stage caused by environmental and strategic factors. These, in turn, determine variation in ultimate wheat yield.

* Taken from "The Use of the Earth Resources Technology Satellite for Crop Production Forecasts", Draft Final Report of the Task Force on Agricultural Forecasting, Goddard Space Flight Center. July, 1974

Figure 1 depicts the interrelationships between the various elements which determine wheat yield. The final yield will be determined by both growth factors and by factors which cause crop abandonment (i.e., failure to harvest the crop).

Growth Factors

The factors that affect the growth of wheat can be divided into those which are determined by environmental factors and those that are related to strategy options available to farmers. The environmental influence consists of a number of factors including soil characteristics, temperature, moisture, light, wind and carbon dioxide. Each of these will now by briefly discussed:

Soil

Soil is a physical medium for plant growth and provides moisture and nutrients to crops. On the other hand, it harbors insects and diseases which can attack plants. The physical quality of soils which are measured by such items as texture, permeability, available water capacity, liquid limit, the plasticity index, density, acidity-alkalinity reaction, and chemical properties (e.g., organic carbon percentage, electrical conductivity, calcium carbonate equivalent etc.) can impede or facilitate the movement of water and certain nutrients such as nitrate and sulfate ions. Because of thier complexity, many of the properties of soils and their interactions with plants have not been quantified. However, it is known that the above-mentioned factors affect most of the stages of plant growth and ultimate yield.

Temperature

Air and soil temperatures significantly affect wheat at various stages of plant growth. Seeds will not germinate if the soil temperature is below $40^{\circ}$-$45^{\circ}$ F. Cooler temperatures usually cause slower growth. The maturities of various plants are determined largely by degree-days.

Moisture

Moisture is the most commonly discussed environmental factor in the literature. The amount of soil moisture at seeding time, the seasonality, frequency and duration of rainfall during the season as well as the total seasonal amount all significantly affect plant development. During the growing season, plant roots take moisture from the soil and transpire much of it back to the atmosphere through the leaves. When soil moisture falls below the wilting point for that soil, the plant becomes moisture deficient and further development is retarded. Decreased yield or plant death could follow. Water accumulating on the surface of the soil can delay planting

or drown or retard the growth of already planted seeds. Heavy rains on gorwing plants can also cause lodging. Lodging can cause plant maturity to be delayed, takes longer to combine-harvest, and can result in the sprouting of kernels that are in contact with the ground.

Light

Light is the catalyst necessary for the conversion of carbon dioxide and water into sugars, protein and ultimately, yield. Latitude and intensity of sunlight are the primary factors. Latitude affects day-length and both short-wave (solar) and long-wave (terrestrial) radiation are correlated with cloud cover. Rates of photosynthesis depend upon the receipt of visible light and rates of transpiration are affected by the net exchange of radiation by the crop canopy.

There is little man can do, at the present time, to control day-length. However, wheat growers can modify the amount that strikes each leaf plant by adjusting seeding rate, distance between plant rows and distance between plants and by breeding new seed varieties with nearly upright leaves in order to minimize shading and maximize the amount of leaf area exposed to sunlight.

Wind

The major effect of this variable is in causing lodging of wheat plants. This could delay ripening and cause problems in harvesting.

Carbon Dioxide

This gas is needed by plants to carry on photosynthesis. Experiments have shown that increasing the atmosphere's concentration of this gas above normal levels increases dry matter significantly. Thus, the composition of the atmosphere will affect wheat yields.

Some methods man could use to modify these environmental factors include:

a. Irrigation is used to augment natural precipitation. The importance of the proper amount of soil moisture both before seeding and during growth has been discussed above.

b. Fertilization - commercial fertilizers supplement soil nutrients in more than half the wheat fields. The dryer the area, the less fertilizer is used. A defficiency of each of 12 essential mineral elements required for plant growth results in a specific change in color and/or shape of the plant. In general, partial lack of a nutrient

causes a plant's leaves to turn some shade of yellow and results in a shorter plant with lower yield.

c. Planting practices include depth of planting, plant spacing and date of planting. Farmers adjust the depth of planting according to soil moisture and temperature. As a general rule, the cooler and moister the soil the closer to the surface the seed is placed in order to provide maximum yield.

Plant spacing affects time of covering the ground, weed incidence, available moisture supply and the amount of leaf area exposed to sunlight and ultimately yield. Research and farmer experience have provided management with the knowledge to consider these factors with a view toward obtaining the highest possible yields.

Generally, the earlier the date of planting of spring wheat, the higher the expected yield. However, this is constrained by soil temperature and moisture and the probable amount of danger from frost for the emergence plants. Planting of winter wheat will generally wait for an adequate level of soil moisture and consider the danger of Hessian fly.

Crop pattern alterations prevent water and nutrient supplies of the soil from being depleted. For example, summer fallowing is carried on in order to store up the years rainfall and accumulate nitrates.

d. Herbicides, insecticides and pesticides are used to control weeds, insects and diseases. Weeds, which can diminish plant population and cause water deficiency can be controlled via herbicides and are less of a problem than diseases and insects which can cause decreased yields or complete crop loss.

e. New seed varieties are used to take advantage of genetic differences among plants. These genetic differences account for differences in the way in which different plants react to environmental factors. Thus, seed breeders are continually developing varieties with varying characteristics of yield potential,

disease resistance, insect resistence, plant height, stalk strength, length of growing season, drought resistance leaf conformation, root conformation and winter hardiness.

As will be seen below, accounting for all these factors simultaneously present a serious problem in any analysis of the causes in variability of crop yields.

## CROP ABANDONMENT FACTORS

Given that one could perfectly model the growth factors, it is still necessary to consider those factors which might lead the farmer to fail to harvest the crop. These can be patterned into natural factors which cause the crop to fail and economic factors which influence the farmers. These factors include a) drought which, although it is at least partially accounted for in consideration of precipitation deserves mention here since it is such a serious problem in some parts of the world, b) wind, hail, winterkill and crop disease which are generally difficult to forecast and not included as explanatory variables in any of the models consulted in this study, c) Insect damage which might be mitigated by the use of pesticides. Note that the environmental failure effects produce significant reductions in the theoretical yield produced by existing models and that the occurrence of these events are potentially detectable from space. Thus, a dramatic improvement in yield prediction could be realized by including these factors in an overall yield model.

The economic impact on crop abandonment is relatively straightforward but is not considered in the yield models discovered during the literature search. The current price of the crop, the cost of harvesting the crop and the government support in the form of crop insurance combine to provide trade-off decisions for the farmer. Planting of winter wheat for forage and/or soil protection with the intention of plowing it under in the spring is a fairly wide spread practice which if unaccounted for could lead to serious bias in estimated yield. In recent times the dramatic increases in wheat price have in some cases led to a harvesting of crops which were originally planted for forage purposes. Thus, if forage is considered in a model, a potential for misspecification in the other direction exists.

## ANALYTICAL APPROACHES USED IN PREVIOUS STUDIES

The documents reviewed at this writing were written for a variety of audiences, on a variety of topics, used different techniques of analysis

and contained differing attitudes and assumptions toward crop yield forecasting. Some yield forecasting models were built for the purpose of estimating the effects of variation of a single policy variable such as irrigation. Other models are concerned with determining the relative effects of several different variables that are known to affect crop yield and thereby understand the structure of the causative factors leading to crop yield. Still other studies estimate a model for the primary purpose of predicting yield. The great majority of the models studied are concerned more with effects of individual factors and policy determination than they are with forecasting.

The techniques used in previous studies include:

a. Regression analysis of local, state and national data

b. Regression analysis of visual quantification of crop conditions for specific localities.

c. Observations of crops under controlled environment

d. USDA surveys of farmers

e. Parametric time-series analysis

f. Estimation of formal production functions.

## REGRESSION ANALYSIS

Regression analysis is the technique used most frequently in previous studies. In this section we will discuss the general types of regression studies encountered in the literature review and the difficulties encountered in these studies which account for so many unsuccessful attempts at forecasting crop yield. A more thorough background for this discussion appears in the reviews of the literature in Appendix F.

### The Nature of Previous Regression Studies

The theory behind most existing models for yield prediction appears to be that air composition and soil fertility exhibit little variation from year to year by comparison with the considerable fluctuations in air temperature and water supply. Positive or negative genetic factors and crop abandonment factors are rarely explicitly considered.

Most of the earlier studies related wheat yields on a local or state basis to environmental conditions such as inches of precipitation or average temperature of critical months. One basic problem in these models is their inability to account for technological change, especially more recent breakthroughs. A typical way of handling this is to use a time trend to represent technological change. This assumes some sort of systematic embodiment of technology.

Another basic problem with some of these models is their use of seasonal and even monthly averages of some of these variables. A number of subsequent phenological and field studies have shown that there is a gradual change of the effect of weather variables on crop yield development throughout the growing season. R. A. Fisher (1924), developed a statistical technique for analyzing the daily effect of rainfall at any time during the growing season. This technique has since been used and modified by a number of studies, especially those involving rainfall as the most critical explanatory variable. The technique involves the estimation of a function of rainfall as a polynomial function of a biometeorological time variable. A similar approach is illustrated by Baier (1973).

As indicated above, there is considerable interaction of causative factors. For example, the use of fertilizer might increase the response of the crop to additional soil moisture or precipitation.

For some meteorological variables their interacting effects have been partially captured by the development of new weather parameters which can be derived from standard climatological data and are related to the way in whic plants and soil conditions react to them. Examples of this are such relatively new concepts as potential evapotranspiration, heat units and soil moisture budgeting. For example, Mack and Ferguson (1968) developed a moisture stress index for a wheat crop using the modulated soil moisture budget developed by Holmes and Robertson in an earlier study. This index is expressed as the difference between potential evapotranspiration and actual evapotranspiration and is found to correlate more closely with wheat yields than other water-related variables tested, such as seasonal precipitation. Nix and Fitzpatrick (1969) develop a crop water stress index which accounted for a greater proportion provided the best statistical results. However, it is possible that poor data reporting systems in Turkey might have made disaggregated data more vulnerable to errors. Williams (1970) estimated yields for each of the crop districts in the Canadian prairies and extrapolated the results for each province and for the Canadian prairies as a whole based upon acreage values, and similarities of environmental conditions. Although the national estimates appear accurate some district and provincial totals were underestimated while others were overestimated thereby compensating each other. Probably, if the errors of the individual local estimates were random, an aggregation of many local estimates would result in a lower standard error for the national total than for the local estimate. However, because of the factors mentioned above, this would require different equations for each local area.

## VISUAL QUANTIFICATION OF PLANT DEVELOPMENT

This technique was developed by Professor J. R. Haun of Clemson University, Clemson, South Carolina. A technique was developed whereby

daily observations of wheat developed was recorded as an index (based upon the rate of development of leaves and other plant parts). This was regressed against age, cumulative development and environmental factors and various lags, transformations and cross products. The observations were made on five wheat plantings in 1966 in Dickerson, North Dakota and the predictive equation was tested using 1967 data. The actual and predicted estimates appear in close agreement. However, some systematic bias is evident. In a paper due to be published this month, the author will demonstrate the use of this model in predictions of yields.

The application of this model to national totals would require extensive gathering of morphological data throughout the growing season.

Chirkov (1973) reports that the Russians have had considerable success in forecasting wheat yields by observing physical characteristics of plant development. For example, for dark soils, the factors described as influencing wheat yield predictions in order of primary importance are number of stems in the spring, phase of emergence of the stalk, number of ear bearing stems in the flowering phase. A secondary factor is the height of winter wheat plants starting from the flowering phase and a tertiary factor is the supply of available moisture in the soil layer from 0-100 cm during the ten days following the resumption of growth in the spring.

A confidence factor of 80% for prediction of the yield of winter wheat is claimed using only moisture supply, number of stems per $m^2$ in the spring or in the phase of emergence of the stalk and, for a forecast prepared in the flowering phase, the number of stems with an ear and the height of the plants. Inclusion of secoadary factors is said to increase the confidence factor to 90 percent.

It is stated without backup that equations have been developed which forecast the yield of winter wheat with great confidence for individual fields, oblasts, regions, republics and for the country as a whole.

OBSERVATIONS OF CROPS UNDER CONTROLLED ENVIRONMENT

Many studies in which plants are grown under controlled conditions are referenced in the literature and several have already been reviewed at this writing. These include wheat grown in greenhouses or on small plots in which almost all factors are held constant except the particular one the experimenter is interested in. The studies that have already been reviewed in this effort include those investigating the effects on yield of changes in soil moisture, different types of herbicides, nitrogen fertilizers, ethral and supplemental irrigation. These studies are generally useful in enumerating factors which affect wheat yield, but are of too limited a purpose to be used to eliminate national crop yields.

## USDA SURVEY TECHNIQUES

A few documents discuss the use of surveys in the U. S. and Australia to forecast crop yield at different times during the growing season. Understanding this technique generally involves two parts: a description of the data collection techniques and a description of the forecasting techniques.

In the U. S., information is collected by mail surveys, telephone contacts, personal interview and observations in selected fields from producers,, feeders, grain elevator operators, and exporters. This information includes acreage intended for planting, planted, intended for harvest and harvested, expected yields and production, inventories,, employment and wages. The results of these surveys are checked for consistency against information collected for the Agricultural Census conducted every five years and other relevant data.

For supplemental information " an objective yield survey is performed in which trained enumerators visit 17,000 sample plots in a sample of fields during the growing season to obtain quantitative data of such factors as number of plants per plot, plant spacings, number of wheat heads and spikelets,, stage of development, final yield and harvesting loss. This information is gathered monthly.

The annual cycle of crop projections begins with a report on farmers intentions to plant. This report is based upon data gathered in the February surveys and is published in March.

The second major survey in early June, when most crops are in the ground, is combined with the June Enumerative Survey and published in the July Crop Report along with estimated production during the forecast season of August through November. An acreage update survey is conducted each July to determine changes that need to be made in June data. This first update appears in the August Crop Report. A third survey effort in the Fall measures acreage actually harvested.

The system for estimating yields relies on a "graphic regression method" which relates reported crop conditions to a forecast of yield. Crop reporters estimate the probable average yield in their localities and the averages of these forecasts are translated into yield forecasts by the Crop Reporting Board by means of regression charts which relate historical "true" yeilds to reported probable yields. In some states, a regression equation is used to forecast yield per acre as a function of a) reported condition of crop (reported yield per acre), b) precipitation for specified months prior to date of forecast, c) precipitation for specified months after date of forecasts and e) time.

Gunnelson, Dobson and Pamperin (1972) examined the accuracy of more than 1,100 USDA crop production forecasts for barley, corn, oats, potatoes, soybeans, spring wheat and winter wheat for the period 1929-1970. He found that USDA forecasts generally exhibit desireable properties based upon his criteria. Unsatisfactory first forecasts were divided almost equally between those which exhibited turning point errors and those which correctly indicated the direction of change but which erred significantly in magnitude First and second revised forecasts showed improvement over the first forecast lowest percentage of satisfactory revisions were found for Winter wheat (59.5 and 52.4 percent for first and second revisions respectively). Although the revised forecasts tended to be successful, they tended to undercompensate for the error in the previous estimate.

In general the accuracy of first forecasts seem to have shown moderate improvement between 1929 and 1970; that of the first revisions remained relatively constant; and that of the second revisions appears to have improved.

Although this study revealed no serious inadequacies in crop forecasts, the analysis identified a few persistent inaccuracies in the forecasts. Specifically USDA tends to:

a. Underestimate crop size

b. Underestimate the size of changes in production from year to year and

c. Undercompensate for errors in previous forecasts when developing revisions.

While USDA crop forecasts exhibit desireable characteristics when appraised by these criteria it is possible that the levels of some of the forecasting errors exhibited may create planning problems for farmers and marketing firms.

PARAMETRIC TIME SERIES ANALYSIS

This technique is based upon two assumptions regarding the factors affecting yield. First, it is assumed that the major factor affecting yields - weather - is difficult to forecast and second, the embodiment of technological change is highly correlated through time. Because of this, an attempt is not made to identify the underlying structural relationships and national average crop yield data is used for identifying and estimating the autoregressive process. The results showed poor forecast accuracy. This appears understandable since from qualitative information we know that yield variation around the time trend is substantial.

## ESTIMATION OF PRODUCTION FUNCTIONS

Studies which estimate production functions so as to compare factor input are of interest in aiding our understanding of the production process but are of limited use in forecasting crop yields.

## SPECIFIC MODELS OF INTEREST

This section discusses the specific models found in the literature to have relevance to crop yield forecasting. Although most of these models are not meant to be used specifically as a forecasting tool they can be adapted for this function and they provide valuable information which can be used to construct such a model. The information provided in the published and unpublished literature is inconsistent with some models described in more detail than others. The time and resources available in this study did not in most cases, allow us to gather data beyond the published literature.

In general, most of the models reviewed in this study would probably not provide as accurate a forecast as does the USDA system for national wheat crop forecasting. This is due to a number of factors. First, these models have not been successfully extrapolated to national totals. This is because they are either estimated from very local data, use very broad assumptions or require quite complex information networks. Second, genetic factors and crop abandonment factors are rarely considered explicitly. Comparisons with local USDA forecasts were generally not performed.

An accurate validation of a forecasting model should include forecasts made beyond or before the sample period for which it was estimated as well as a full description of statistical tests and of the behavior of the model during the sample period. Such a description should include a discussion of mean error as well as extreme errors and a full explanation of how well the model predicted turning points. In view of these criteria, discussion of validation of these models is slight or nonexistent.

Variables related to water use by plants appear to be the most significant variables in these models. These include soil moisture, moisture stress, potential and actual evapotranspiration and combinations of these. Furthermore, the effects of these variables change with the age of the plant.

We will now briefly discuss a few of these models which appear to offer some merit in deriving a forecasting model. Table 1 has been prepared as a handy summary of the properties of these models:

### Weather and Canadian Prairie Wheat Production

This study by G. D. V. Williams (1960) reports on the use of regression techniques to analyze wheat production. The dependent variable

was wheat yields in various regions in Canada. Explanatory variables were:

a. Precipitation conserved in the 21-month summerflow period prior to May 1st of that year

b. Precipitation for May, June and July (three variables) and,

c. Estimated potential evapotranspiration for May, June and July (three variables)

d. Various combinations and powers of the above although these variables are listed, the actual equations used were not presented in the document reviewed. It is stated that there were a number of different equations estimated for different time periods from 7 to 14 years between 1952 and 1967.

District crop yield estimates are then extrapolated to a total for the Canadian prairies according to a weighting system using acreage values.

Using equations based on data prior to 1960, estimates of wheat yields were made for the period 1960 to 1967 based on precipitation and PE data available before the end of July, June and May, respectively. For this period the extrapolations appeared to catch turning points and direction quite well although they did not reflect year to year differences very closely. Although 1961 was an unusually poor year, the estimate was close. This indicates that in practice, if weather-based estimates were being made for the current year, the equations could be developed from, say, the preceding ten years rather than an equation that was estimated for a period ending several years earlier. Estimates made on data available at the end of June would probably be very close to those at the end of July. However, those performed at the end of May are less accurate.

Although national estimates appear accurate, some district or provincial totals were underestimated while others were overestimated thereby compensating each other.

Wheat Production In Turkey

A study published by the U. S. Department of Agriculture in 1970 reports on regressions of wheat yields against weather conditions during different parts of the growing season, mechanization and fertilizer use over the period 1948-1968. Weather conditions for all 12 months of the year were tested for significant correlation with wheat yields as was a mechanization variable. The best equation was:

$$Y = 883.9 - 2.03 X_5 + 11.15 X_{12} + 13 X_{13}$$

$$t = \qquad 2.93 \qquad 4.31 \qquad 3.04$$

$$R^2 = 0.82 \qquad SD = 104.3$$

where

$X_5$ = January - February aridity index for Ankara

$X_{12}$ = May - June aridity index for Ankara

$X_{13}$ = Fertilizer consumed in 1,000 metric tons

The standard deviation is about nine percent of 1968 yields values. When the equation was used to predict yields beyond the sample period (1948-1968), the error was less than five percent for 1969 and 1970. The error for 1971 was not reported in the paper. However, it is cautioned that since the standard deviation is nine percent, this sort of accuracy is not likely to hold further into the future. The model would have to be updated periodically since the methods and patterns of wheat production in Turkey are changing rapidly.

The Thompson Model

L. M. Thompson (1969) estimated a number of regression equations of time trends and weather variables on wheat yields for six states (North and South Dakota, Kansas, Oklahoma, Indiana, and Illinois). Weather variables included state averages of precipitation, rainfall and temperature for various months throughout the year. There has been some criticism of the use of state averages of weather variables since wheat is not evenly distributed throughout the state. However, there is some "tendency for favorable or unfavorable conditions from year to year to be fairly widespread."

The six equations estimated are presented in the original review in Appendix 1. Coefficients of determination ranged from 0.80 to 0.92 and standard errors ranged from about 9-12 percent of 1968 yield.

The only hint of an attempt at validation in this paper is a graphical comparison of the model's estimates with those of USDA.

The Baier Model (1973)

This model incorporates several new features which take advantage of recent developments in the understanding of agrometeorological interrelations. Instead of using rainfall data, the model uses potential evapotranspiration (PE) and soil moisture (SM) as independent variables. In

addition the concept of biological time (BT) (rate of development toward maturity) is introduced.

It is assumed that the yield response of a crop to these variables changes gradually over the season and that the daily weighting of each variable can be adequately fitted by a fourth-order polynomial as a function of bio-meteorological time. These functions are estimated by an iterative regression process. These estimates are then used as explanatory variables in a multiplicative regression model. This technique is further explained in the appendix.

The equations derived are not presented in the paper, but the variables used are maximum temperature, minimum temperature and soil moisture as functions of time. The best coefficient of determination was 0.79. The model was not used for forecasting beyond the time period or lattitude in the sample.

Although the methodology appears to show ptoential for accounting for daily changes in plant response to environment, the present model cannot be used successfully as a forecasting tool since it has not been tested, the data is quite dated (1953-1962) and the results have not been extrapolated to national totals.

Proprietary Commercial Models

The documents consulted in this study consisted primarily of those that have been published through journal articles, universities and domestic and foreign governmental agricultural services. However, in our various telephone conversations with experts in this field around the country we have become aware that there are a number of models in existence constructed by private firms for commercial purposes. The exact structure and estimation techniques used are said to be proprietary and therefore these models are not generally available for detailed review. However, a general description of a model available through the Development Planning and Research Associates, Inc., (Manhatan, Kansas) is provided here:

The DPRA model is claimed to have overcome many of the short-comings of the regression models discussed above by considering simultaneously much detailed information regarding the phenology and production of wheat (and other crops) into a detailed structural model of the plant growth process. This model includes all of the crop growth factors mentioned above (including both environmental factors such as temperature, soil moisture, solar radiation, soil characteristics and man made factors such as irrigation, fertilizer, weed and insect control, time of planting, depth of planting and

rate of planting) as well as genetic factors such as maturity ratings of various varieties of plants in various different climates.

The model has been used primarily for two purposes. The first is to advise farmers on policy such as irrigation, fertilizer and cropping patterns. The second use for this model is in forecasting yield. DPRA claims to have a much greater degree of accuracy in this use, than the presently available USDA forecasts. These forecasts are available throughout the season beginning shortly after planting. DPRA also states that although present forecasts are regularly performed only on a field and regional basis the model can be expanded to national and worldwide levels with only a minimum effort.

The model might be useful for any group wishing an additional dimension with which to check forecasts made through other means.

## CONCLUSION ON STATUS OF AGROMET MODELING

We have seen that yield variation is caused by many growth factors (environmental and genetic) and by crop abandonment factors (environmental and economic). None of the yield forecasting models reviewed in this study included crop abandonment factors. The nature of the specific effect on yields of the growth factors are extremely complex in that a) their affects vary with different stages of the crop growth cycle, b) their effects are often lagged in complex distributions over time and c) they interact with each other in complex ways many of which are undefined.

Because of these complex factors, regression analysis, which has been widely used in numerous studies has been unable to capture the underlying structural relationships of yield determination. The number of variables that can be successfully used in a regression equation is far fewer than the number of variables that affect crop yield. Furthermore, most of the previous regression models were estimated for local or state areas and cannot be satisfactorily extrapolated to national and world totals without a massive data gathering effort.

Variables related to water use and temperature for certain critical periods in the plant growth cycle are consistently the most important variables in the studies consulted. In recent years, new ways of measuring these variables (potential and actual evapotranspiration, moisture stress, soil moisture budgeting and biological time) have shown promise of possibly improving the predictive ability of regression equations. However, these models still account for only 70 to 90 percent of the variation in yield and have large standard errors of estimate.

Based on these large standard errors, on the results of the few models that were examined for predictive accuracy and on the fact that these models are generally valid only for a specific local area, it appears that none of these models can predict national crop yields as accurately as the USDA survey-judgmental system. This conclusion does not preclude the use of some of these models as additional input to a judgmental process.

Recent advances in models which incorporate plant observations with soil moisture data appear to hold some promise for accurate yield predictions since the entire history of both environmental and genetic effects is presumably contained in the current state of the plant. In some cases, these visual observations are related to plant density and are therefore potentially observable from space.

A realistic procedure for synoptic predictions of wheat yield might be the development of ground truth in selected sites coupled with sample survey techniques to develop region yield/acre estimates. This would be followed by intensive monitoring of these sites (remote and relayed in situ) by satellite coupled with satellite estimates of variations in harvested acreage resulting from crop abandonment factors.

Although these models have only limited use in forecasting compared to the methods used by USDA, they are valuable in providing much information regarding yield-environment interactions and in that recent advances provide hope for increased accuracy sometime in the future. In areas of the World where extensive data gathering networks are nonexistent, agricultural forecasting models which rely on satellite data inputs might be able to improve upon present forecasts.

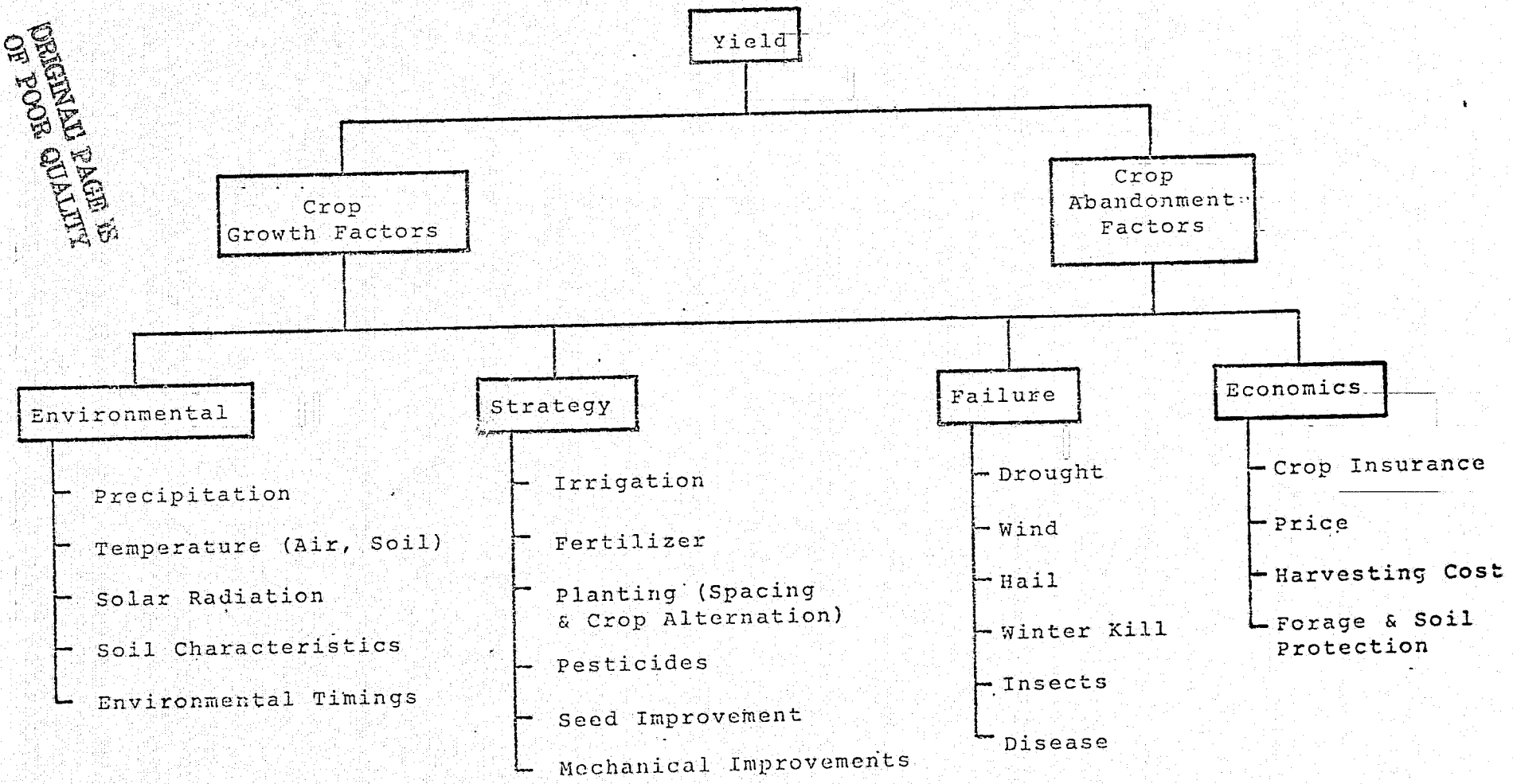
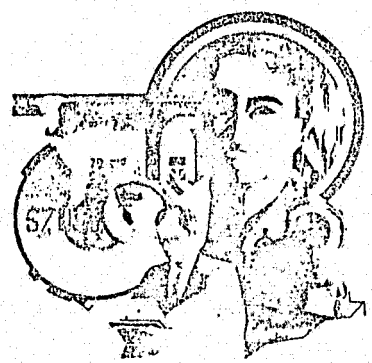



Figure A.1 Wheat Yield Factors



# Chapter 2

# SAMPLING METHODOLOGY AND ESTIMATION

## INTRODUCTION

Although the Statistical Reporting Service conducts some of its surveys by virtually complete enumeration of certain parts of the population, most are based on samples drawn from the population. With the use of modern techniques, sampling is not only less costly in time and money than a census, but also can produce more reliable results.

The Service uses a great variety of sampling techniques to produce current agricultural statistics about crops, livestock, prices, and other information relating to the agricultural economy. Significant advances in methods used have been made in recent years, particularly with the emphasis on probability sampling technology, although nonprobability sampling retains an important place in the work of the Statistical Reporting Service.

This chapter provides a description of the common sampling procedures (frame construction, sample selection, analysis, and estimation) currently used and describes some of the research activities under way to improve the quality of agricultural statistics.

## THE SAMPLING FRAME AND SAMPLE SELECTION

A basic consideration in any sample survey is the sampling frame, which is an aggregate of units or elements from which a sample can be selected. From data collected in the sample, inferences may be made about all the elements in the frame. These elements collectively form the survey population, which may or may not be the same as the target population, which is the total universe of elements about which information is desired. From SRS surveys, estimates must be made for the target population.

The type and quality of sampling frames have much influence in determining sample design and overall survey methods. The frames used by SRS are of two basic types—the list frame and the area frame.

### List Frame Sampling

Sampling from list frames has for many years played a prominent role in the collection of data for agricultural statistics. A list frame is a list of elements presumably all from the population about which inferences are to be made, along with appropriate identifying data. Lists of farm operators, including names and addresses, are used for many of the surveys conducted by SRS and are well suited for the collection of information by mail. The low cost of data collection from a list sample is one of the principal advantages of this method. Another advantage is the ease with which supplementary information for classifying the units can be included as part of the frame. This allows the use of efficient stratified sample designs.

The main disadvantage of the list frame is the inability to compile "complete" lists; that is, lists that represent all of the current units, such as farms, livestockmen, or processors—such units are continually changing. For example, a list of farm operators soon becomes outdated because new operators enter the activity, others leave the farm, some expand operations or lease land to others, or there are other changes within the operations themselves.

Since probability sampling requires that all units of the population be represented, list sampling had few applications for probability surveys until relatively recent developments permitted selection from two or more frames that cover the population. Applications of such multiple-frame sampling are discussed later in this chapter.

Prior to the application of probability sampling by SRS during the early 1960's, nonprobability mail surveys were the principal means of collecting data for current agricultural statistics. This method is still used as an important data collection technique for many commodities, but usually requires supplemental survey information.

In using nonprobability mail samples, the shortcomings must be recognized. First, lists of potential respondents are not complete frames and, while still useful, some lists tend to be selective as well. Second, there is no assurance that respondents who voluntarily complete and return a questionnaire are typical or representative of those who fail to do so. The second limitation can be overcome with followup interviews of at least a sample of nonrespondents. However, this is usually not practical, considering the limitations imposed by the frame, and nullifies the principal advantage of nonprobability mail surveys—low cost.

Despite the biases inherent in mail samples, surveys of this type with sufficient response provide consistent indications from survey to survey. Appropriate methods of estimation are used to remove biases from the estimates insofar as possible.

## Area Frame Sampling

In 1954, SRS began investigating the use of area frame sampling. A program was developed and expanded to include the 48 conterminous States by 1967 in a system of surveys for obtaining information on crops, livestock, and other agricultural items. Today area frame sampling is an integral part of the SRS estimating program.

In area frame sampling the frame consists of an aggregation of identifiable units of land (segments) which may be sampled. For SRS purposes, characteristics concerned with agriculture must then be associated with these sample segments. There are three different concepts that are useful in associating agricultural activities with the area frame. These are the closed segment, the open segment, and the weighted segment.

The closed segment associates the agriculture with the segment itself; it includes all that is inside the segment boundaries and excludes all that is not. In the open segment, all activities of farms with headquarters located inside the segment boundaries are associated with the segment regardless of whether the activity itself is inside or outside the segment boundaries. In the weighted segment, all agriculture associated with a farm, any part of which lies within the segment, is attributed to the segment in proportion to the fraction of the farm acreage that is inside the segment.

For characteristics such as crop acreages which are directly associated with land, the closed segment has proved to be clearly superior in sampling efficiency. But data concerning the economics of the farming enterprise, for example, can be more easily associated with the farm headquarters and do not lend themselves to the closed segment. The weighted segment is used to gain efficiency by reducing variability caused by specialized and widely differing sizes of farms.

A unique attribute of the area frame is that it is a complete sampling frame. All desired agricultural activities are represented when every unit of land area has been given some positive probability of being selected during the sampling process. Furthermore, it does not suffer the same kind of deterioration through time as does a list frame.

The area frame lends itself well to enumerative general-purpose surveys. It is not suited to mail surveys, since names and address of persons living or operating within the segment boundaries are generally not known in advance. The area frame is not efficient for special-purpose surveys or surveys of highly specialized farming activities, because the lack of supplementary information precludes the segregation of farming enterprises of a particular class.

Two basic types of area frames are in use by SRS for general-purpose surveys. The first is the frame developed for the Master Sample of Agriculture, which was constructed in the early 1940's at Iowa State University with the cooperation of USDA and the Bureau of the Census. The Master Sample was designed for sampling characteristics associated with farms. The frame consists of

county maps upon which minor civil divisions and frame units containing a specified number of sampling units have been delineated. Each sampling unit contained about four farms. SRS experience suggested that segments half the size of those of the Master Sample were more efficient for general-purpose surveys, and these units are being used. Crop reporting districts are used to impose geographic stratification on the frame. Typically, States contain about nine crop reporting districts. Within these districts the agriculture is fairly homogeneous. Allocation of segments to crop reporting districts is about proportional to the square root of value of products sold.

The Master Sample frame was available for use from the beginning of SRS area frame sampling. However, it was soon apparent from pilot work in the Mountain States that stratification of land according to use was essential. Consequently, the second type of area frame used by SRS is the land use frame, in which all land prior to sampling is first classified according to use. The stratification is based on extent and type of farming and can be described in four broad categories: (1) Intensively cultivated areas where a significant portion of the land is under cultivation, (2) extensive agricultural areas used primarily for grazing and producing livestock, (3) highly developed land found in cities and industrial areas, and (4) nonagricultural land, such as parks and other recreational areas. In addition to land use stratification, geographic stratification is frequently used to separate differing agricultural areas.

Segments are of a predetermined size, with segment counts associated with each area delineated on maps according to size of area. Segments typically are about 1 square mile in intensively cultivated areas, several square miles and larger in the more open farming areas, and about one-tenth square mile in city and residential areas. The number of segments sampled from each stratum is determined by reviewing optimum allocations for major commodities and choosing a compromise for general-purpose sampling.

Land use frames are currently being developed State by State as needs indicate and as time and resources permit. States still using the Master Sample frame are in the north central, south central, and south Atlantic regions, where differences of land use practices are less apparent.

Segment selection has generally followed a systematic-sample approach where the frame listing is arrayed geographically. Recently, interpenetrating sample designs have been used. Interpenetrating designs utilize several smaller independent samples, and have more sample flexibility and advantages in computing sample variation. They also fit well with a sample rotation scheme. Typically, 20 percent of the SRS segments are rotated annually to relieve respondent burden.

All selected segments are visited annually about June 1 for the June enumerative survey to ascertain planted crop acreages and inventories of hogs and cattle, and to classify operations for purposes of subsampling for subsequent surveys. All separate land operating arrangements are delineated within the segments and are referred to as "tracts." To control sampling errors, the area sample is supplemented with a small list frame sample of known large livestock operations, this being a limited form of multiple-frame sampling.

Sampling for several subsequent area frame surveys uses the June information for classifying tracts. The classifications made are utilized as strata for second-stage sampling. Tracts are then subsampled from each stratum at varying rates, according to their information potential. The December enumerative survey is the largest survey of this type and focuses on fall-seeded crops and livestock inventories. A large portion of the tracts with wheat and livestock in June are selected. Nonagricultural tracts are sampled very lightly.

### Multiple-Frame Sampling

A method rapidly gaining importance and use in SRS surveys is multiple-frame sampling. As the name implies, this technique includes the use of more than one sampling frame. For SRS needs, this means a list frame and an area frame.

Theory for multiple-frame sampling was developed only as recently as the early 1960's. Research under the leadership of Dr. H. O. Hartley[1] was supported by SRS at Iowa State University. Concepts of multiple-frame sampling are basically those of probability sampling concerning repre-

[1] Dr. Hartley is currently Director, Institute of Statistics, Texas A&M University.

sentation, known probabilities, and randomness of selection. In addition, two criteria need to be considered: (1) Every element of the population must belong to at least one of the sampling frames, and (2) it must be possible to identify for each selected unit to which frames, if any, it belongs other than the one from which it was selected. The use of a complete area frame satisfies the first consideration. The second is more difficult operationally, requiring the proper classification of each tract operator as to whether he is also included in the list frame.

Multiple-frame sampling has some distinct advantages for SRS, particularly for items such as livestock, specialized crops, and economic data. These items are poorly correlated with land alone and are inefficiently estimated by the area frame. In multiple-frame sampling, most of the data for the population of interest can be collected more efficiently through the list frame. Some of the data can be collected by mail. Also, it is usually possible to develop and incorporate in the list frame some index of size for units that is used in stratification. The area frame measures list incompleteness. In this way, the two frames complement each other.

The State Statistical Offices have principal responsibilities for developing list sampling frames of farmers and ranchers for multiple-frame surveys. A variety of list sources is used, including State farm census, assessor's records, Agricultural Stabilization and Conservation Service (ASCS) lists, brand lists, and lists maintained by State governments for inspection or control purposes. More specialized lists are often combined with a basic list to improve list coverage. Lists vary greatly in quality and usefulness and require considerable effort to prepare before use in sampling.

Often the list has to be converted into computer-readable form. Units which are duplicated must be removed and the indexes of size of operation may have to be obtained from other sources. Special large mail surveys are sometimes conducted for the sole purpose of classifying farms by type and size. County and local officials of ASCS, the Extension Service, and other USDA agencies have provided valuable assistance in list development efforts.

After initial list development, maintenance and updating are continual tasks. Without such efforts, lists deteriorate rapidly and soon lose their advantage in sampling efficiency.

## ESTIMATION METHODS

After a survey is designed, the sample selected, and data collected, the data must be edited for consistency and then summarized. From these survey results the statistician must prepare the estimates. The computations and procedures for translating survey data into estimates involve technical considerations. Usually more than one method is available, but the choices are largely specified by survey design and there are distinct differences between deriving estimates from nonprobability surveys and from surveys which follow the concepts of probability theory.

### Nonprobability Surveys

In developing current estimates from nonprobability mail surveys, estimating procedures must recognize potential biases in the survey results. The procedures used generally depend on past relationships of survey data to final estimates. It is assumed that these same relationships are continuing, but periodic checks must be made to verify this assumption and to true up the estimates. Check data are obtained from a variety of sources, but generally are in the form of records of marketings or census enumerations. Information from the U.S. census of agriculture and from annual farm censuses conducted in some States has commonly been used for this purpose.

Many factors affect the reliability of estimates derived from nonprobability surveys. First, it is necessary to evaluate the accuracy of the check data used to establish true values. Errors in these data will result in errors in the relationships derived for past years. There is always the possibility of error in assuming that past relationships of survey data to final estimates will continue. Comparability of survey data must be maintained for the period in which relationships are derived. If survey indications for past surveys are based on selective data, indications used to make the current estimate must be subject to the same kind of selectivity for best results. Therefore, consideration of comparability should be given to the list samples, the sampling procedure and distribution, and the survey response.

### Survey indications

Direct-expansion indications are not possible with nonprobability surveys because of the inability to associate known probabilities with the data collected. Therefore, most survey indications are relationships estimated from the survey data which can be applied to some assumed known base. A brief description of some of the commonly used nonprobability survey indications follow.

*Ratio to land:* Relations of an item to total land in farms can be estimated from survey data. Used primarily for crops, the sample total acreage for a specified crop is divided by the sample total farmland acreage. This provides a measure of the proportion of farmland acreages used for individual crops. The relations of any two items on the questionnaire can be estimated in this manner.

*Ratio to base:* This indication is similar to the above but the control variable, such as capacity of feedlots or grain storages, is known in advance and is part of the sampling frame. The ratio estimated from the sample totals can be expanded by the known base totals for the population.

*Average per farm:* Averages per farm estimated from survey data are used to estimate livestock. These averages can be associated with estimates of farm numbers. Averages obtained from mail surveys can be quite biased because of widely varied farm sizes, which may not be properly represented among survey respondents.

*Matched reports:* Estimates of survey-to-survey changes can be made by matching "identical farm" reports from two successive surveys. This indication has commonly been called the "current/current" ratio. Indications are developed by applying survey changes to the previous estimates. Care must be taken in the matching process to assure that the reporting units are comparable between surveys. The procedure does not permit new operating units to be included in the tabulations.

A variation of this procedure is the "current/historical" indication, which also measures change from some previous period, but data for the prior period is collected on the current questionnaire. For example, a farmer would be asked to report his previous year's acreage of each crop along with current year's acreage. The advantage is that all reports can be used for tabulation and no matching is required, but it has been found that the data reported by farmers for the preceding year are often subject to error because of memory bias or other reasons.

*Yield indications:* Mail surveys have retained much of their usefulness for estimating and forecasting crop yields. Perhaps one reason is that yields do not vary greatly by size of farm. At harvest, actual yields can be derived by obtaining harvested acreage and comparable production data. Indications for forecasting yields are based on reports of condition or probable yield. Reported condition consists of evaluations by growers and crop reporters of the size of the current crop expressed as a percentage of a hypothetical full or normal crop. Expected or probable yield is likewise a subjective judgment of crop prospects, but is expressed directly as yield per acre.

### Data interpretation

The assumptions that must be made to prepare estimates from nonprobability survey indications are factors that limit survey reliability. Several methods most frequently used for minimizing or interpreting the inherent biases should be mentioned.

*Weighted averages:* A procedure for minimizing response biases is to use geographic or size group stratification in summarizing the data. Known or estimated weights are used to weight stratum averages up to State estimates. The effect of a poor distribution in sample response is minimized, providing respondents have characteristics similar to others in the same stratum. For example, crop yields would normally be expected to be more alike within a crop reporting district than within an entire State. Average yields from the survey are computed at the level of the crop reporting district and weighted to a State average yield, using district estimates of crop acreages for weights. Size group stratification is used similarly.

*Charts:* Most nonprobability survey data are interpreted in some way through charts which pictorially describe past relations of survey data to final estimates. The most common of these is the simple regression chart, where the relations are plotted, using the horizontal axis for locating the magnitude of past survey indications and the vertical axis for corresponding estimates. The statistician prepares the estimate by determining the best-fit location on the graph corresponding

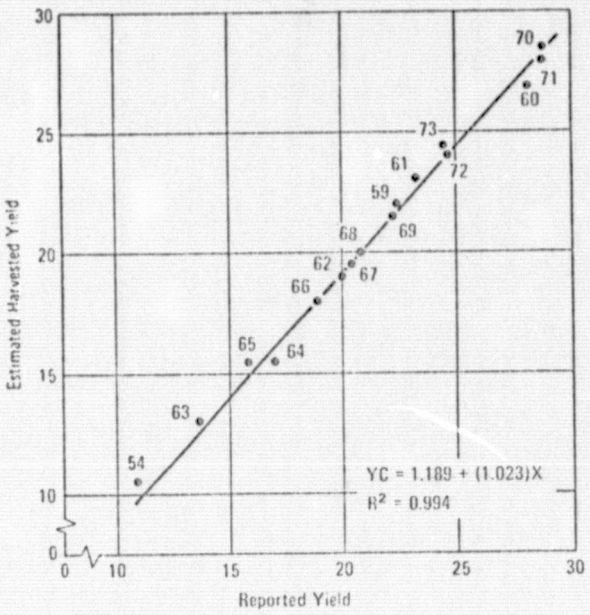



Figure 1.—Example of a regression chart used to estimate a State's winter wheat yield.

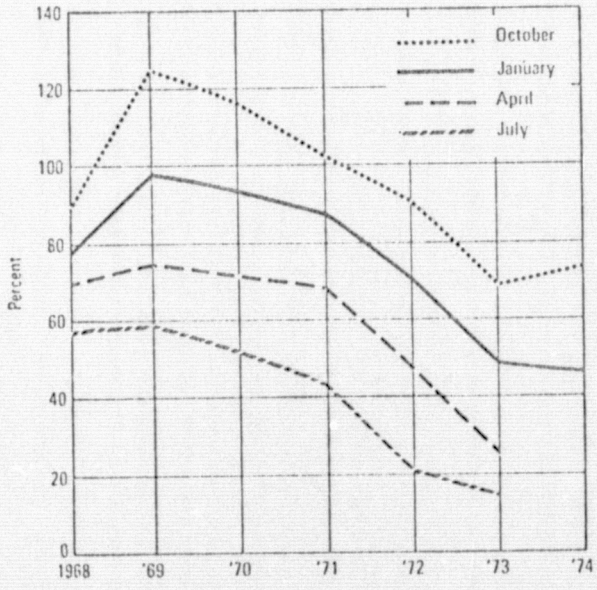



Figure 2.—Example of a time-series chart used in estimating a State's stocks of wheat on farms.

to the current survey indication. The graph interpretation is frequently done visually, although the linear regression line is usually computed and plotted to assist interpretation. Points on the graph are identified by year so that recent year relations can be given more influence if desired.

Time-series charts are used for some commodities. The horizontal axis is used for the sequential plotting of time, and the levels of indications and estimates are indicated on the vertical axis. Indications and the corresponding estimates are distinguished by different types of lines drawn to show respective year-to-year changes. Current estimates are set with the available knowledge of these past relations between the level of estimates and survey indications.

Trend is an important consideration for some estimates, particularly in developing crop yield forecasts. A time-series chart in addition to a regression chart is sometimes used for this purpose. The regression chart is used to present the usual survey-estimate relationship. Deviations from the regression line are plotted on the time-series chart. These deviations plotted sequentially illustrate the effect of time and allow a projection to be made. Another method uses time as a second variable for developing a multiple-regression indication. In this way an allowance for trend is incorporated into the indication. Additional variables, such as precipitation, are occasionally used in developing the multiple-regression indication.

## Probability Surveys

Estimates can be made from probability surveys without dependence on prior survey relations or benchmark data. With known probabilities, raw data are expanded into unbiased estimates of current agricultural activities. Also, sampling errors are computed that provide the statistician with a tool for evaluating the reliability of estimates generated. Sampling errors not only provide measures of precision, but the sources of sample variation are useful in optimizing sample designs and allocations. The quality of statistics derived from probability survey data usually justifies their higher costs.

Basic considerations for survey reliability are sampling frame, survey design, and sample size. Each is important in maintaining sampling errors at acceptable levels, although constraints on sam-

ple size are frequently imposed by budget limitations. Measures of nonsampling errors are rarely available. Much effort is made to minimize potential nonsampling errors through survey training programs, questionnaire design and testing, providing precise survey procedures, and utilizing comprehensive editing systems.

### *Enumerative survey*

In SRS "enumerative survey" refers to area frame sample surveys in which data are collected by personal interview. The basic estimator used for area frame survey data is the unbiased direct expansion. Raw survey data from each segment are expanded by the reciprocal of the probability of selection. Estimates are generally computed at the stratum level for analysis purposes, but inferences from enumerative survey data are seldom made below the State level, because of relatively large sampling errors. Segments are the primary sampling units, hence tract data must be summed to the segment level. Sampling errors are then determined from the variation between segments.

Ratios and ratio estimators are also utilized with data from area frame surveys. These estimates are particularly helpful in evaluating changes from survey to survey. For the June enumerative survey, ratios are computed by comparing current survey data with previous-year data for identical segments. Ratios are computed at each level of summary, hence biases inherent in ratio estimates are minimized. In expanding previous and current matched data, consideration is given to the fraction of total sampling units that are comparable. If 80 percent of the segments in a stratum are identical (following a 20-percent annual rotation scheme), all expanded matched data would be divided by an additional factor of 0.8. This allows for variations in the rotation scheme. The ratio estimate is derived by applying the ratio to the direct expansion estimate from the previous year's survey. Estimated sampling errors take into account the correlation or covariance of the matched data.

A third estimator is derived from a ratio to land area. This estimator is efficient for major crops and other items that are highly correlated with land area. The actual area of each segment is measured from a scaled aerial photograph. The relation of each item to the measured area is calculated and this ratio is applied to the total base land area at the State level. All concepts of ratios and ratio estimates apply; however, the base or total land area is assumed to be known without error.

Somewhat more difficult are the theoretical concepts associated with subsequent area frame surveys in which all June tracts are first classified into strata and then subsampled. Although it is a two-stage sample design, the second stage of sampling is not confined to primary sampling units, as it is in cluster sampling. Instead, the second stage of selection is among all tracts classified according to predetermined criteria using the June information. With this sampling scheme it is quite likely that some segments will have no tracts selected in the sample. Unbiased direct-expansion estimates can still be generated by associating the probabilities of selection (probabilities at the first stage of selection multiplied by probabilities at the second stage) with the data for each tract sampled. The difficulty arises in computing sampling errors. The concept assumes a product estimator where the factors are a population estimate for total number of tracts within each classification and an estimated average tract value for tracts within each classification. The variance component associated with estimating the number of tracts is computed from the June enumerative survey, whereas the component for between-tract variation must come from current survey data.

Ratio estimators are also used for surveys based on subsamples of June area tracts. The ratios are computed by relating current data to June data. The June enumerative survey direct-expansion estimate becomes the base for computing a ratio estimate. These estimates are particularly useful for the July acreage update survey where correlations are very high between actual planted acreages and those reported during the June enumerative survey (which in some cases are intended plantings).

### *Multiple-frame survey*

The general estimation model for multiple-frame surveys based on a list and area sampling frame is:

$$X = X_a + pX_{al} + qX'_{al}$$

where $X_a$ = the estimated total for the portion of the population included only in the area frame;

$X_{al}$ = the estimated total for the population included in both frames, computed from the area sample;

$X'_{al}$ = the estimated total for the population included in both frames, computed from the list sample;

and $p + q = 1$

Since $X_{al}$ and $X'_{al}$ are two independent estimates of the same population (overlap domain), any values for the weights p and q which sum to 1 will provide unbiased estimates. Optimum weights will be inversely proportional to the variances associated with each estimate. In practice, weights are predetermined, utilizing information from prior surveys. The value of q is usually large and is associated with the greater efficiency of the list frame. For livestock surveys, values of $p = 0$ and $q = 1$ are used. This equation is often referred to as a "screening" estimator. In the variance computation, $X_a$ and $X_{al}$ are considered nonindependent components of the estimating equation.

Little use has been made of ratios and ratio estimates in multiple-frame sampling. Direct-expansion estimates have proven to be efficient and allow complete flexibility in developing the sampling plan for each survey.

#### *Objective yield survey*

Objective yield surveys provide crop yield information for estimates or forecasts based directly on counts, measurements, and weights of the crop made from small plots in a probability selection of sample fields. When a crop is mature and ready for harvest, yield can be estimated by harvesting and weighing production from these plots of known size and expanding to a yield per acre. This method of preharvest sampling to estimate yields is often referred to as "crop cutting." Similar procedures are used for tree crops, but yield is computed in terms of production per tree and observations are usually made on sampled limbs. For a mature crop, estimating yield becomes primarily a sampling problem. Theoretically, samples can be designed to produce estimates of yield with any desired degree of precision.

The same sampling considerations are important for objective surveys used in forecasting yields. In addition, early-season plant characteristics must be identified which can be used to predict yield at maturity. A forecast model (often a regression equation) has to be developed that describes the relations between the prediction variables and the final outcomes. For all crops, it is usually helpful to analyze yield in terms of two components: Number of fruits and weight per fruit. Reliable forecasts of number of mature fruits are readily possible, since most plants set fruit at a fairly early stage of maturity. Identifying useful plant characteristics and predicting weight per fruit is more difficult, since growth of the fruit typically continues until maturity.

An additional factor of yield which must be taken into account for SRS estimates is harvesting loss. Biological (gross) yields can be estimated from preharvest objective sampling but these estimates overstate the production that is actually hauled from fields and can enter marketing channels. To estimate net yield, special post-harvest surveys are conducted to measure all production remaining in fields after harvest. These losses, which are measured by gleaning small sample plots immediately following harvest, must be subtracted from gross yield.

#### Field crops:

Concepts and general methodology used in objective surveys for forecasting and estimating yields are similar for all field crops. Sample fields are selected from fields identified during the June enumerative survey as having the crop of interest. A systematic sampling scheme is used for selection, following a geographical arrangement of fields. Self-weighting samples are achieved by assigning probabilities of selection which are proportional to expanded field acreages. This facilitates summarization and has proven to be efficient for estimating purposes. Observations are made on two randomly selected plots (units) in each of the selected fields.

Objective yield surveys are planned to coincide with the publication of production forecasts and estimates in the monthly Crop Production report. During the first survey month, crop maturity will vary considerably by area of the country. Appro-

priate counts, measurements, and other observations are made for each sample that will be used in the forecast models. Plant characteristics used as prediction variables change as maturity progresses. At an early stage, for example, a count of plants may be the only data available, but it is valuable in forecasting the number of mature fruits. If no characteristics are available to predict weight per fruit, historical averages will be used for the sample. As the crop matures, other variables become important. Actual fruit counts are used, and weights and measurements of the immature fruits are often useful in predicting final weight per fruit. Simple linear- and multiple-regression models are most often used to describe past relations between the prediction variables and the final observations at maturity. Typically, relations observed over the preceding 3-year period are used in current forecast equations. Forecasts of gross production are computed for each sample. Plots for most crops include two adjacent rows of predetermined length. Measurements are made to determine row spacing so that conversions can easily be made to yield per acre. An adjustment is made for expected harvesting losses, based on past averages. Individual sample yields are averaged to arrive at State estimates. Sampling errors are based on variation between sample yields.

As the season progresses and crops mature, the individual sample yields provide data for estimates rather than forecasts. Final preharvest observations are made as near harvest as practicable. Similarly, for best results it is desirable to do the postharvest work immediately following farmer harvest. When the information is available, actual harvesting losses are used in computing net yields.

### Tree crops:

Sampling frames used for selecting blocks (fields) of trees have been developed by various means. In some cases, nearly complete lists of growers, classified by size of operation, have been made available through trade or marketing associations. Area frames have been constructed by identifying blocks of trees on aerial photographs. Stratification according to age of tree reduces sampling variability in some applications. In addition to its uses in sampling, the frame usually becomes the basis for estimating the population of trees.

Blocks of trees are sampled with probabilities proportional to the number of trees or acres, which results in a self-weighting sample. Counts are usually made on two to four trees per block. A random method is used for selecting a "pivot" tree with additional count trees selected nearby. This cluster reduces counting time within the block. The random-path method is commonly used for selecting count limbs on a tree. Beginning at the base and proceeding up the tree, a random selection is made at each point of branching until a count limb of suitable size is obtained. Probabilities proportional to the cross-sectional areas of the limbs are usually used in the selection process to gain sampling efficiency. An alternative to the random-path method is to select a primary limb as described, but map out the remaining branches into suitable count limb sections. A random choice of one or more of these sections can then be used for counting purposes. On mature trees, 5 to 10 percent of the tree is usually counted. The probabilities associated with each stage of selection must be used in expanding the limb counts to an estimate of fruit per tree.

Once fruit is set, forecasting becomes the task of projecting drop and growth. Most droppage occurs immediately following bloom, after which the fruit counts become relatively stable. Predicting weight of mature fruit is done by relating immature size or weights to final weights. Drop and growth patterns observed in past surveys are a requirement for the current forecasts.

Periodic surveys are used to update the projections of fruit drop and growth until harvest. An allowance must be made for fruit remaining after harvest, particularly if mechanical harvesting equipment is to be used. Since blocks are the primary sampling units, sampling errors of estimated production per tree are computed from variation between blocks.

## PREPARATION OF ESTIMATES

Forecasts and estimates represent the combined effort of both the State Statistical Offices (SSO's) and the Washington, D.C., offices. Most sample data are collected, edited, summarized, and analyzed in the SSO's. State statisticians prepare the initial forecasts or estimates for their States and transmit them with supporting data and comments

to the Crop Reporting Board in Washington for review. An explanation of unusual local conditions or other pertinent information affecting an estimate is given in the statistician's comments.

In Washington, the State data are summarized nationally for each item. Estimates recommended by the State statisticians are reviewed by commodity specialists of the Crop Reporting Board. The reviewers have all the survey information that was available to statisticians in the States and can evaluate the data at the national and regional levels. For many commodities, State survey indications are summed for the U.S. level and a national estimate is set first. These procedures permit the use of check data and other survey information available at the national level. For example, some of the probability survey data are extremely valuable at the national and regional levels, but are more limited in value for State estimates because of relatively large sampling errors.

For all major commodities, including livestock species and crops identified as speculative, members of a formal Crop Reporting Board convene to review and adopt the official estimates. Each member makes an independent interpretation of all available data and recommends an estimate. The Chairman of the Board reviews these recommendations and reconciles differences of opinion.

## RESEARCH

SRS continually conducts research aimed improving the quality of its services to the public. The principal areas of study are briefly described below.

### Sampling-Frame Construction and Maintenance

Through the past several years research and operational experience have resulted in an evolution of area frame construction. Most States now have a land area sampling frame based on stratification of land according to agricultural use. A recent advent to the basic design has been the use of interpenetrating sampling to select units from the frame. Within a land use stratum a set of independent samples are selected, using a random method. Interpenetrating sampling facilitates an orderly rotation plan of sampling units for enumeration. Other advantages are that a replication can be used as an independent estimating sample for special purposes, and the land use stratum variance may be computed quite easily by using the replicated means or totals.

Research in land area sampling-frame construction centers on fine tuning, or introducing greater efficiency in the methodology. Current investigations cover optimum stratification and segment size; ways to improve accuracy and quality-control measures; and exploration of new frame materials, such as high-altitude or satellite photographs. Since the land area sampling frame is the only complete sampling frame, SRS must maintain and improve the efficiency of its use, even though SRS relies heavily on the sophisticated application of list files as a partially complete frame for estimation.

A second area of research is in developing name list files suitable for use in multiple-frame sampling. A major problem associated with constructing such a file is identifying duplication of names within the file. The process of identifying duplication using computer technology is called "record linkage." Specifically, record linkage brings together two or more separately recorded pieces of information concerning the name of a particular individual or operation. Tasks within the overall heading of record linkage include data manipulation (the process by which unlike records are restructured to make them more comparable without changing the basic information) and information coding (the process of removing variations of alpha or numeric information by substituting a common code system). By performing these two steps, the similarity of records has been increased without changing their information content. Once these two processes are completed, it must be decided if individual records are linked with other records. Probabilities are used by a model to create the likelihood of link or nonlink, and a hypothesis test is used in deciding if two records are indeed the same. Finally, a method is developed by which information gained about name records may be retained so that the process of identifying unique list name sampling units improves over time through survey use.

### Nonsampling Error

Research on nonsampling errors is directed at

the survey as an instrument to measure certain items of interest, such as crop acreage or numbers of livestock.

Nonsampling errors are to be distinguished from the sampling error, which arises from the use of a sample rather than the entire universe of elements to be studied. All other types of error are called "nonsampling errors," a term often loosely considered as synonymous with "response errors" and "measurement errors." Nonsampling errors are not necessarily related to the size of the sample, as are sampling errors. They may arise from errors of measurement, since any measuring instrument will vary in its ability to measure precisely the item of interest. A survey is subject to many sources of nonsampling errors: The frame may be unsatisfactory, sample selection may be biased, questionnaire design may be deficient, improper information may be recorded, mistakes may be made in processing the data, and data may be missing because of lack of response, etc.

Unlike sampling errors, nonsampling errors present considerable difficulty in the estimation of the variability that may be associated with them. It may be possible to measure some particular component of such errors, but there may still exist some unknown components. As a result, there has been little practical work done in the area of estimating nonsampling errors. More progress has been made in identifying sources of nonsampling errors.

Identifying the sources of nonsampling errors is the first step in developing procedures to remove them. Analysis of survey data and comparison of results of independent surveys measuring the same items may indicate sources of nonsampling errors. Sometimes such analyses or comparisons indicate that nonsampling errors are present, but do not identify the sources. If this occurs, an alternative is to reinterview by an independent method that is considered to be more accurate. This can be done with a subsample of survey respondents. It is assumed that the reinterviewing team is a more accurate measuring instrument, because better interviewers are used and the questionnaire is structured in greater detail to reveal the correct values if they are not obtainable by a direct question.

After sources of nonsampling errors are identified, it is necessary to develop procedures to measure the degree to which they affect the items of interest. One procedure is to use replicated sampling to build into a survey an experimental comparison of several different measuring processes, providing the measuring devices do not have the same type of systematic errors. Another procedure is to assign replications to interviewers to determine the variability in survey data that is attributable to the interviewers when it is not a systematic error. The idea is to make part of the survey a controlled experiment with precautions, such as randomization, that are typical of good experimentation.

Refusals are responsible for part of the nonsampling errors due to nonresponse. Procedures are developed and tested not only to reduce the number of refusals, but also to provide estimates of those that remain refusals.

### Remote Sensing

"Remote sensing" means measuring an object or phenomenon from a distance, whether by photography or other radiometric technique using microwave instruments, spectroradiometers, multispectral scanners, etc. These measurements are of electromagnetic energy which is emitted, scattered, or reflected by the objects observed. Different objects return different kinds and amounts of energy. Remote sensing utilizes these detectable differences to identify ground objects or phenomena from the air or from space.

Crop identification and acreage measurement have been recognized as potential applications of remote sensing. An ideal approach might be to make acreage estimates from sensor information every 24 hours, but the data-handling problem and the lack of an all-weather sensor system makes this impossible except in special situations. Consequently, other ways have to be found to use remote-sensing data.

Several possible approaches are: (1) double sampling or multistage sampling, (2) multiple-frame sampling, or (3) using space imagery as an area frame on which broad land use classifications have been done. This land use classification would then be used in designing a stratified sample. Or space imagery could be used as a frame from which one could select a subsample of aircraft flight strips and, within flight strips, select area

segments. These area segments could then be photographed at a larger scale or enumerated on the ground. This is a multistage sample using several different kinds of information.

Likewise, space imagery of a county or State could be classified according to the crops of interest. From this classification one would select a sample (or use an existing sample) of area segments and collect the necessary information about these areas on the ground. This is a double-sampling technique in which the space information is the large sample, and ground survey provides the more detailed information. If the correlation between the ground information and the space data is high, substantial gains can be realized in making crop estimates for the total area.

Space imagery may also provide more efficient estimates by providing supplementary data. For example, it may be possible to classify crops by frame units in the present area frame. This would mean that if one were interested in corn, he could select the sample from frame units with probability proportional to the acreage classified as corn. If the correlation between the classified corn acreage and the actual acreage was high, gains in estimation using ratio and regression techniques could be realized.

Until an all-weather satellite is developed, an estimating technique must be developed that can be used where satellite coverage is incomplete. One solution is to use multiple-frame estimating techniques, such as using the space imagery to estimate the cloud-free area, and the aerial photographs and ground enumeration estimates for the area covered by clouds on the space imagery. Then, by proper weighting, all three data sources are combined to obtain an estimate for the total area.

Remote sensing has some potential in livestock estimation, particularly in hard-to-get-to areas or in areas of nonresponse. At present, this approach is limited to aerial photography with sufficient resolution and to areas where livestock occupy open areas, or areas with limited vegetation.

## Yield Forecasting and Estimation

Research directed toward the development of objective methods of estimating and forecasting yields is conducted for a wide variety of crops. The estimation of crop yields at harvest and forecasting of yields yet to be realized are two distinct phases of the research effort. For most crops, the development of methods of estimating harvesting losses constitutes an additional phase.

Crop yield estimation is based on the observation of plant and fruit characteristics just prior to harvest, at harvest, or soon after harvest is completed. Research in estimating biological yield, harvested yield, and harvest losses involves developing methods which rely on statistical sampling and estimation theory. For purposes of efficient sampling and estimation, it is often useful to treat yield as the product of components such as weight or size per fruit, fruit per plant, and plants per acre.

Forecasting of yields involves predicting what has not yet happened. Methods of forecasting the final yield while a crop is still immature are obviously more difficult to develop than estimation procedures at harvest. Crop yields are the culmination of many factors. These factors are generally associated with the plant, its location, weather, and production practices. The timing and interaction of weather factors and the extremely complex interactions of all important factors make their direct use in predicting final yields extremely difficult. Fortunately, observations of the immature crop can be made which are often useful in predicting the resulting yield. Crops in an immature stage of development are a reflection of the collective and interacting effects of these factors over a portion of the growing season. Inasmuch as these same factors also constitute a primary influence on the mature crop, observations made at an immature stage provide a good basis for yield forecasts.

To develop successful methods of forecasting yields, it is necessary to discover specific plant characteristics which are useful predictors of yield. A comprehensive understanding of the fruiting behavior of a crop is the essential first step in the development of the predictive models. Forecast models designed to relate these characteristics to yield or its components may be based upon knowledge, verified by experimental studies, about plant growth and development during the season and time-related growth patterns. This knowledge may be acquired primarily through agricultural

research. Special investigative surveys are made to fill gaps in previous research and to adapt the models to current practices. In addition to models which rely on the repeatability of plant growth, and patterns adjusted for current fruit development, regression models based on the stability of parameters between years are in use. These models often incorporate the developmental stage of the plant and its fruit in order to utilize unique model parameters for individual maturity categories by States or agricultural regions.

Forecasting crop yields also requires efficient estimation of variables used in the models which have been developed. Sampling and estimation theory is utilized to achieve this efficiency. Since sampling considerations are usually sufficiently compatible for the predictive variables and estimates of final yield, the same sampling design can be used for obtaining both immature and mature plant and fruit observations. Thus relations between observations at various stages of maturity may be studied in great detail at the common elementary unit level or at other levels in a hierarchical sampling design. ■

# APPENDIX C

## THE REMOTE SENSING OF BARE FIELDS FOR CROP ACREAGE ESTIMATION

If automatic processing of LANDSAT digital data for full-scale crop surveys is to become a reality, the solution of crop classification problems by use of spectral signatures of growing crops is required. In particular, considerable effort must be expended on the technical problems of: (i) signature extension (ii) supervised and unsupervised classification "learning" algorithms (iii) spectral signature analog areas, etc., all applied to the crops in various stages of their growth cycle.

On the other hand, principal investigator Stanley A. Morain has done a successful Kansas 10-county winter wheat study relying on the correct classification of freshly plowed "wheat" fields - implying the intention to plant wheat - with subsequent adjustments due to the growth and harvestability of the actual wheat.* His method required visual interpretation of the imagery, and thus may not be found suitable for adaptation to automatic processing. Concerning the difference in approach between Morain's study and others, note the following points:

*Kansas Environmental And Resource Study: A Great Plains Model; Extraction of Agricultural Statistics from ERTS-1 Data of Kansas, S.A. Morain, Type III Final Report under contract NAS 5-21822, Task 4, February 1974.

(1) It is relatively "easy" to discriminate freshly plowed fields from the same fields covered with stubble from the last harvest, or fallow, i.e., containing some plant cover or containing worthless crops left to rot or used for forage.

(2) Crop calendars, throughout the world, are well known and documented (in the statistical sense).* This does not give one certainty as to what will be planted at a particular point in time in a specified field; but it provides a high probability that a known crop will be there, or in the case of crop rotation, that one out of two or three crops will be there.

(3) The intelligent use of crop calendars, as by Morain, should provide an excellent database together with LANDSAT data from which to construct the initial acreage estimates. These must be corrected later for losses (very occasionally also gains) due to hail, flooding, late frost, insect infestation, blight and farmer's decisions not to harvest. These points are discussed in more detail in the notes at the end of this appendix.

Thus, the initial LANDSAT acreage estimates based on plowed fields correspond to USDA/SRS "planting intentions," but of course are much more nearly objective. Furthermore, they can be done on a near census-type approach, rather than using a tiny probability sample with relatively large sampling errors.

---

*Agricultural Atlas

The crop calendars, which should be very detailed, contain the essential information for classifying with LANDSAT a large fraction of agricultural acreage in the U.S. (also in other countries with similar agricultural practices) at the time of planting* - or shortly before. This provides a good estimate of planting intentions acreage. The fields should be catalogued, for later information retrieval, so that the growth of a healthy crop can be verified, or in cases of severe crop stress the acreage can be accordingly reduced. Eventually, the LANDSAT system may also be capable of detecting crop condition sufficiently accurately to allow for the measurement of yield by using intertemporal data on already classified fields, thus supplying a complete remote sensing system for obtaining crop production estimates. In the meanwhile, it is important that the acreage estimation be done as well as possible with LANDSAT.

Investigators who are working with remote sensing of growing crops as compared to bare fields appear to be attempting to solve a much more difficult task, i.e., of resolving the intricate spatial and temporal differences in spectral signatures of growing crops.

Granted there is a need for this effort in attempting to develop a complete crop production measurement system with LANDSAT.

---

* Which varies both by crops and by country, and within country, by latitude and geography as well.

But pioneers of the crop survey applications effort might have a better chance of succeeding in the near future if they would start with the acreage estimation, using the available information as to what most likely will be planted in the freshly plowed fields from knowledge of the local planting times.

Writing in the Type III Final Report (1974) of "Kansas Environmental and Resource Study: A Great Plains Model," Morain stated:

> "The results presented here demonstrate that a simple method for winter wheat identification may be developed given an adequate prior knowledge of local environment and crop cycle. The method appears to be applicable to other crops if suitable distinct crop cycle events may be defined. Knowledge of the local environment is critical if the interpretation is to be successfully conducted. Components of the local environment data set can be taken directly from the ERTS-1 imagery (Williams and Coiner, 1973) but other components are best developed at the local level. Furthermore, surface observations for a small number of fields from each environmental area would be a necessity. The necessity for (1) surface observation, (2) knowledge of the local environment, (3) knowledge of the local crop cycles, and (4) the modest amount of equipment and training required to perform these interpretations make this method suitable for implementation at the local (county) level."

## NOTES

1. Acreage nearly always <u>decreases</u> from planting time onwards through the growing season due to the simple fact that crops suffering various kinds of stress may be (i) plowed under (ii) left to rot in the field (iii) destroyed completely by hail or floods. Nevertheless, occassionally there are increases due to replanting with another crop. e.g., a corn crop

is planted on May 1, damaged by flooding on May 15 and subsequently plowed under in late May. The same field is then replanted with soybeans on June 1 resulting in a net loss of corn acreage, but a net gain of soybean acreage.

2. Plowing a field under may be done at various points in the agricultural cycle:

- post-harvest and pre-planting, usually in Spring,
- pre-harvest by farmer's decision relating to expected profits.

In the latter case, various reasons for plow-under exist:

- to give nutrients to the soil,
- to allow for a second crop to be planted.

Whenever plow-under occurs in preparation for planting a field, it is a normal part of agricultural practice.

3. The economic decision not to harvest crops already planted happens very infrequently in poor countries. In rich countries it may be done for reasons of crop stress, poor yield or low price expectations and is usually accompanied by plow-under. However, in the case of 1974 flooding in the Mississippi Valley, the crops were left to rot in the fields. If there is a decision to replant a field, it will usually allow a small time window for LANDSAT to observe the change in the field - perhaps a week at most.

4. In tropical countries, the ability to grow more than one crop per year makes crop classification by remote

sensing more difficult. This remark applies particularly to India. Nevertheless, there may be a chance to observe the plowing between crops.

# APPENDIX D

## SAMPLING PROBLEMS IN REMOTE SENSING CROP SURVEY APPLICATIONS

### The LACIE Program's Sample Design

Suppose that the sample consist of M area segments (e.g., 5 x 6 mile rectangular areas, as in LACIE) selected according to a stratified sampling plan for the crop in question. Each segment contains N pixels of which a fraction, $f_c$, are cloud covered. Because the supervised classification procedure uses previously designated training fields with each segment, it is necessary to obtain a certain minimum level of cloud-free pixels. Segments which have too large a value of $f_c$ will be rejected (approximately $f_c > .20$). Wheat acreage is estimated by a weighted sum of the "wheat" pixels in the segments which pass the cloud cover test.

$$W = a \sum_{i=1}^{M} \delta_i w_i \sum_{j-1} \lambda_{ij} Z_{ij} \Big/ \left(\sum_{j=i} \lambda_{ij}/N\right)$$

where: a = area of 1 pixel

$\delta_i$ = 1 if $i^{th}$ segment is rejected because of too much cloud cover

$\delta_i$ = 0 if $i^{th}$ segment is accepted.

$\lambda_{ij}$ = 1 if pixel (i,j) is cloud free,

0 otherwise

$w_i$ = sampling weight for $i^{th}$ segment as determined by sampling plan

$Z_{ij}$ = area fraction assigned to wheat in pixel (i,j) - from classification (1 or 0 perhaps?)

Rewriting,

$$W = a \sum_{k=1}^{m} W_{i_k} \sum_{l_k=1}^{n_k} Z_{i_k j_{l_k}} / (1-f_{c,i_k})$$

where $m \leq M$, $n_k \leq N$ and $\{l_k\}_{k=1}$ represents those segments for which $f_{c,i_k} \leq f_o$

and $\{j_{l_k}\}_{l_k=1}^{n_k}$ represents those pixels in the $(i_k)^{th}$ segment which are cloud-free.

There are two problems with this. (1) The subset of the weights in the acceptable (cloud-free) segments does not provide the correct normalization, and (2) The number of segments (m) and the number of cloud-free pixels in each segment ($n_k$) are random variables.

Problem (1) could be attached by artificially forcing the weights to reflect the actual segment selection process. However, this throws the burden onto Problem (2), as the corrected weights will now depend on m and $\{n_k\}_{k=1}^{m}$.

With regard to LACIE, the problem is further complicated by the Group III procedures: for counties which are not represented in the reduced (acceptably cloud-free) sample of segments at all, ratio estimates are concocted using last

year's census figures. While this may be a desirable step for regional or district reporting purposes, it adds no useful information to the national acreage estimate and creates further problems for statistical analysis of the properties of the estimator. The national acreage estimate should be handled in a way that uses the current acreage information from remotely sensed data after cloud cover screening optimally. This purpose is not served by adding in agricultural survey or census data in an ad hoc manner to compensate for missing segments.

Cloud Cover and its Effect on Crop Acreage Estimates

Cloud cover has two effects on the statistical properties of remote sensing estimates of crop acreage: (1) it reduces the available sample at any one time thus causing an increase in variance of the estimate; (2) it may introduce bias into the estimate if cloudiness is correlated with presence or absence of the crop and no adjustment in sample design or estimation procedure is made. At the segment level it is unlikely that the cloud cover distribution is anything but random, and LACIE procedures notably imply this assumption. However, at the district or regional level, it is quite likely that cloud cover distributions will exhibit marked patterns of spatial correlation which leads to the possibility of bias errors in estimating crop acreage. For example, if segments in the state of N. Dakota are frequently covered with clouds or free of clouds simultaneously (if you lose one, you lose them all),

then any peculiarities of wheat culture in that part of the country will be misrepresented in the sample. Although, perhaps, nothing can be done about the missing remote-sensed data per se,* the estimation procedure should be compensated for the effect. One way to do this would be to design the sample - select the segments - with cloud cover as well as wheat culture in mind.

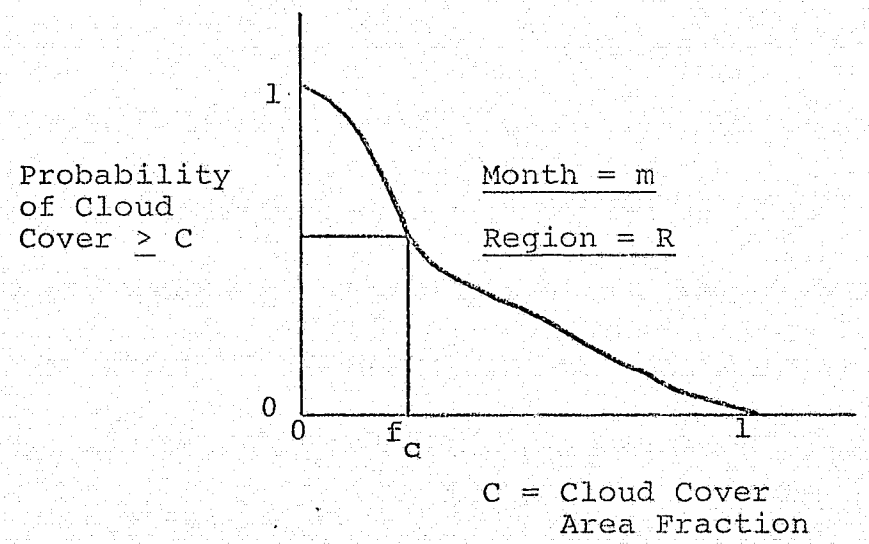



Figure D.1 Cloud Cover Statistics by Weather Region by Month

Suggested Methodology: Obtain cloud cover statistics by weather region by month (see Fig. 1). Determine theoretically a threshold fraction (cloud cover area) $f_c$ for acceptance of a segment.

---

* Consideration could be given to the use of aircraft to fill in gaps in the sample.

This fraction should be as large as possible consistent with the classification procedures. Then stratify the sampling design according to cloud cover in the same way that it would be stratified for wheat growing. These stratifications can be done either in series or in parallel. For instance, the sampling may be done in two stages. First the total list of all segments in the country must be stratified according to the wheat-growing practices. Then select $N_1$ segments from the wheat-growing strata with probability proportional to size (amount of wheat growing in the stratum historically). If $n_j$ is the sample size in the $j^{th}$ stratum, then $N_1 = \sum_{j-1}^{s} n_j$, where s is the number of wheat-growing strata. These $N_1$ segments must then be stratified again by cloud cover probabilities, i.e., the new strata are homogeneous weather regions. From each cloud cover stratum, select some number of segments (p.p.s.)* and form a subsample of size $N_2 < N_1$. Obviously to obtain $N_2 = 640$, it may be necessary to use considerably larger first-stage sample size, $N_1$. If $m_k$ is the sample size in the $k^{th}$ cloud stratum then $N_2 = \sum_{k=1}^{t} m_k$, where t is the number of cloud strata.

The approach outlined above is frequently employed in large surveys. It has the advantage of reducing bias in estimation of the key attributes, while maintaining sampling efficiency.

---

* Probability proportional to size.

## Effect of Spatial Correlations Between Neighboring Segments on Sample Design

The estimates of wheat acreage are not directly affected by spatial correlations between segments (as already mentioned they may be indirectly biased through cloud cover effects); but the confidence intervals are affected as the following analysis shows.

Let $X_i$ = observed no. of "wheat" pixels in $i^{th}$ segment; and $M_i$ = true no. of "wheat" pixels in $i^{th}$ segment.

Let $X = \sum_{i=1}^{M} W_i X_i$ be the wheat acreage estimate where $W_i$ are sampling weights. If the $X_i$ are independent binomial random variables,

$$\text{then var } (X) = \sum_{i=1}^{M} W_i^2 \text{ var } (X_i) = \sum_{i=1}^{M} W_i^2 \sigma_i^2 = \sum_{i=1}^{M} W_1^2 N \left(\frac{M_i}{N}\right)$$

$$\left(1 - \frac{M_i}{N}\right) = \frac{1}{N} \sum_{i=1}^{M} W_1^2 M_i (N - M_i).$$

Now suppose that the $X_i$ are dependent, in a specific pattern indicated by the subscript differences as follows:

$$E[(X_i - M_i)(X_j - M_j)] = \begin{cases} \sigma_1^2 = & \text{if } i = j \\ \sigma_i \sigma_j \rho & \text{if } i - j = 1 \\ 0 & \text{if } i - j > 1 \end{cases}$$

Then $\text{var } (X) = \sum_{i=1}^{M} W_1^2 \sigma_1^2 + \sum_{j=2}^{M} \sigma_{j-1} \sigma_j \rho W_{j-1} W_j$

where $\sigma_1^2$ is the same as before.

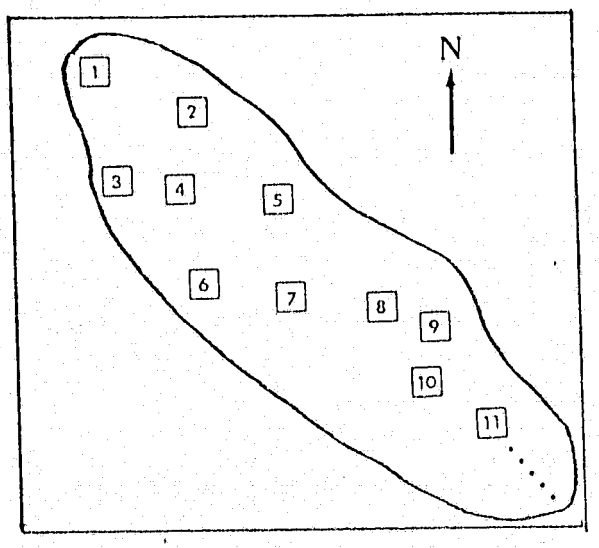



Figure D.2 Numbered Segments in a Wheat "Belt"

The last term is non-negative if $\rho > 0$,* so that the pattern of spatial correlations has caused that much increase in variance of the wheat acreage estimator.

Had this particular spatial correlation pattern been known in advance, even if the size of $\rho$ were unknown, one could have placed a constraint on the sampling plan*: do no select $S_{i-1}$ or $S_{i+1}$ if $S_i$ is selected. For the same size sample this constraint would have increased efficiency (narrower confidence limits) because it would have eliminated the term with $\rho$ in it by causing $W_{j-1} W_j = 0$ for all $j = 1, \ldots M$. The conclusion is that a study of the spatial correlations would generally improve the sampling efficiency.

---

* The occurrence of $\rho < 0$ for spatial phenomena of this type is not plausible: it would involve the implication that wheat acreage is lower in the "neighboring" segment if it is higher in this segment. However for widely separated segments this could occur for economic reasons. The full treatment of this subject would require that advantage be taken of the entire correlation matrix, if known.